REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (*Leave blank*)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
		Meeting Speech

4. TITLE AND SUBTITLE

Sinusoidal Coding

5. FUNDING NUMBERS

C — F19628-95-C-0002
PR —

6. AUTHOR(S)

R.J. McAulay, T.F. Quatieri

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Lincoln Laboratory, MIT
244 Wood Street
Lexington, MA 02173-9108

8. PERFORMING ORGANIZATION REPORT NUMBER

MS-11427

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Department of the Air Force

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

ESC-TR- 96-037

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

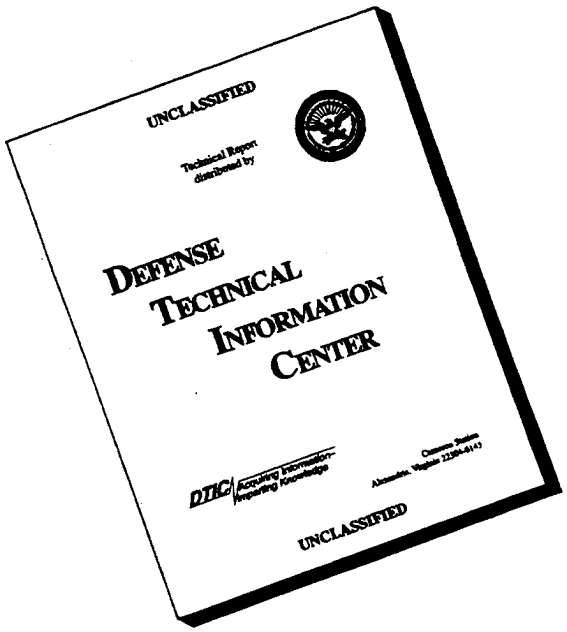
12b. DISTRIBUTION CODE

13. ABSTRACT (*Maximum 200 words*)

14. SUBJECT TERMS	15. NUMBER OF PAGES 53
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unclassified

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

CHAPTER 4

Sinusoidal Coding

R. J. McAulay and T. F. Quatieri

Speech Systems Technology Group
MIT Lincoln Laboratory
244 Wood St., Lexington, MA 02173, USA

Contents

This work was sponsored by the Department of the Air Force

Speech Coding and Synthesis
Edited by W.B. Kleijn and K.K. Paliwal

1. Introduction

One approach to the problem of representing speech signals is to use the speech production model in which speech is viewed as the result of passing a glottal excitation waveform through a time-varying linear filter that models the resonant characteristics of the vocal tract. In many applications it suffices to assume that the glottal excitation can be in one of two possible states corresponding to voiced or unvoiced speech. In attempts to design high-quality speech coders at the mid-band rates, generalizations of the binary excitation model have been developed. One such approach is multipulse [1] which uses more than one pitch pulse to model voiced speech and a possibly random set of pulses to model unvoiced speech. Another is *code excited linear prediction* (CELP) [2] which models the excitation as one of a number of random sequences or "codewords" superimposed on periodic pitch pulses. In this chapter the goal is also to generalize the model for the glottal excitation, but instead of using impulses as in multipulse or random sequences as in CELP, the excitation is assumed to be composed of sinusoidal components of particular amplitudes, frequencies, and phases [3].

A number of other approaches to analysis/synthesis that are based on sine-wave models have been discussed in the literature. The phase vocoder [4] was, perhaps, the first attempt to represent the speech waveform by a set of narrowband functions. A set of fixed bandpass filters is used and one sine wave per filter is assumed to pass within each filter. The frequency deviation of the sine wave from the center frequency of each band is estimated via the phase derivative of the filter output. This frequency deviation is quantized and used in the vocoder synthesis. Portnoff [5] refined the phase vocoder by representing each sine wave component by excitation and vocal tract contributions. The sine-wave frequencies in the model were constrained to be harmonically related. Another refinement of the phase vocoder was performed by Malah [6] who assumed the sine-wave frequencies were harmonic and then made the filter bank pitch-adaptive thus ensuring roughly one sine wave per filter.

The analysis in these systems does not explicitly model and estimate the sine-wave components, but rather views them as outputs of a bank of uniformly-spaced bandpass filters. The synthesis waveform can be viewed as a sum of the modified outputs of this filter bank. Although speech of good quality has reportedly been synthesized using these techniques, the fact that only the phase derivative is coded means that absolute phase information is lost and this leads to degraded, reverberant speech, particularly for low-pitched speakers.

A different approach was taken by Hedelin [7] who proposed a pitch-independent sine-wave model for use in coding the baseband signal for speech compression. The amplitudes and phases of the underlying sine waves were explicitly estimated using Kalman filtering techniques, and each sine-wave phase was defined to be the integral of the associated instantaneous frequency. As in the phase vocoder, absolute phase information is lost. Another sine-wave based speech compression system has been developed by Almeida and Silva [8]. In contrast to Hedelin's approach, their system uses a pitch estimate during voiced speech to establish a harmonic set of sine

waves. The sine-wave phases are computed at harmonic frequencies from the short-time Fourier transform. To compensate for any errors that might be introduced as a result of the harmonic sine-wave representation, a residual waveform is coded along with the underlying sine-wave parameters. To represent unvoiced speech, the model uses a set of narrowband basis functions [9]. Another approach to modeling unvoiced speech in the context of the sine-wave model is to explicitly generate noise via the linear filtering of white noise whenever unvoiced speech components are detected in different bands. This approach, developed by Griffin and Lim [10] as the *multi-band excitation* (MBE) vocoder, uses overlap-add reconstruction for synthesizing unvoiced speech in unvoiced bands. For voiced bands the system uses a "bank of oscillators" which is simply another term for the sine-wave analysis and synthesis scheme described in this chapter. Kleijn [11] has used a version of the sine-wave system in the context of *prototype waveform interpolation* (PWI) to improve the quality of voiced speech synthesis in a CELP coder. More recently, Kleijn and Haagen [12, 13] have used it to define sine-wave frequency tracks along which complementary high- and low-pass filtering operations are done to separate rapidly-varying sine-wave parameters from slowly-varying sine-wave parameters to establish a basis for high-quality speech coding at 2400 b/s.

In this chapter, a sinusoidal model for the speech waveform is derived which leads to an analysis/synthesis technique that is characterized by the amplitudes, frequencies, and phases of the component sine waves. The objective is to demonstrate how a "minimal" parameter set can be derived from this representation, and how this parameter set can be coded for high-quality speech at bit rates from 4.8 kb/s to 2.4 kb/s. Section 2 describes the basic sine-wave analysis/synthesis framework. The general model, which produces high-quality reconstruction for a large class of acoustical sounds including speech, music, and biologics, has too many parameters to be coded at low data rates. By focussing on the speech signal, speech-specific models are developed for the sine-wave frequencies, phases and amplitudes which are more amenable to efficient quantization. In section 3 a sine-wave based pitch estimator is derived and used to replace the sine-wave frequencies by a set of harmonically related sine waves. Section 4 describes a minimum-phase harmonic sine-wave speech model that depends on an envelope fitted to the sine-wave amplitudes and the pitch and voicing parameters. Then in section 5 the amplitude envelope is further constrained to be all-pole and methods for quantizing the parameters of the all-pole model are described for operation in the range from 4.8 kb/s to 2.4 kb/s. Finally, some applications of sine-wave coding to contemporary digital communications problems will be discussed including network capacity optimization, multi-speaker conferencing and the *improved multiband excitation* (IMBE) coder [14].

2. The basic sinusoidal analysis/synthesis system

In this section the sine-wave representation and the corresponding analysis/synthesis system are developed drawing extensively from the authors' ear-

lier published work on this topic [3, 15]. In the analysis stage, the amplitudes, frequencies, and phases of the model are estimated on a frame-by-frame basis, while in the synthesis stage these parameter estimates are interpolated to allow for continuous evolution of the parameters at all the sample points between the frame boundaries. The resulting sine-wave analysis/synthesis system forms the basis for the material presented in the remainder of the chapter.

2.1. The sine-wave speech model

In the speech production model [16], the speech waveform $s(t)$ is assumed to be the output of passing a vocal cord (glottal) excitation waveform through a linear system representing the characteristics of the vocal tract. The excitation function is usually represented as a periodic pulse train during voiced speech, where the spacing between consecutive pulses corresponds to the "pitch" of the speaker, and is represented as a noise-like signal during unvoiced speech. Alternately, the binary voiced/unvoiced excitation model can be replaced by a sum of sine waves, [3, 15]. The motivation for this sine-wave representation is that voiced excitation, when perfectly periodic, can be represented by a Fourier series decomposition in which each harmonic component corresponds to a single sine wave. More generally, the sine waves in the model will be aharmonic which occurrs when periodicity is not exact and when the excitation is unvoiced. Passing this sine-wave representation of the excitation through the time-varying vocal tract results in the sinusoidal representation for the speech waveform, which, on a given analysis frame is described by

$$s(n) = \sum_{\ell=1}^{L} A_\ell \cos(\omega_\ell n + \phi_\ell) \tag{2.1}$$

where A_ℓ and ϕ_ℓ represent the amplitude and phase of each sine-wave component associated with the frequency track ω_ℓ and L is the number of sine waves. The accuracy of this representation is subject to the caveat that the parameters are slowly-varying relative to the duration of the vocal tract system response.

2.2. Estimation of the sinusoidal speech parameters

The problem in analysis/synthesis is to take a speech waveform, extract parameters that represent a quasi-stationary portion of that waveform, and use those parameters or coded versions of them to reconstruct an approximation that is "as close as possible" to the original speech. If the speech waveform is represented by an arbitrary number of sine waves, the unconstrained parameter estimation problem, although easy to solve, leads to results that are not physically meaningful. Consequently, the approach taken here is heuristic and is based on the observation that when the speech is perfectly periodic, the sine-wave parameters correspond to the

harmonic samples of the short-time Fourier transform (STFT). In this case, the model in eq. (2.1) reduces to

$$s(n) = \sum_{\ell=1}^{L} A_\ell cos(n\ell\omega_0 + \phi_\ell) \tag{2.2}$$

in which the sine-wave frequencies are multiples of the fundamental frequency ω_0 and the corresponding amplitudes and phases are given by the harmonic samples of the STFT. If the STFT of $s(n)$ is given by

$$S(\omega) = \sum_{n=-N/2}^{N/2} s(n)exp(-jn\omega) \tag{2.3}$$

then Fourier analysis gives the amplitude estimates as $A_\ell = |S(\ell\omega_0)|$ and the phase estimates as $\phi_\ell = argS(\ell\omega_0)$. Moreover, the magnitude of the STFT (i.e., the periodogram) will have peaks at multiples of ω_0. When the speech is not perfectly voiced, the periodogram will still have a multiplicity of peaks but at frequencies that are not necessarily harmonic and these can be used to identify an underlying sine-wave structure. In this case the sine-wave amplitudes and frequencies correspond to the peaks of the periodogram and the sine-wave phases are computed from the corresponding real and imaginary parts of the STFT.

The above analysis implicitly assumes that the STFT is computed using a rectangular window. Since its poor sidelobe structure will compromise the performance of the estimator, the Hamming window is commonly used to reduce the effect of sidelobe leakage. While this results in a very good sidelobe structure, it does so at the expense of broadening the mainlobes of the periodogram estimator. Therefore, in order to maintain the resolution properties that are needed to justify using the peaks of the periodogram, the window width is made at least two and one-half times the average pitch period. During frames of unvoiced speech, the window is held fixed at the value obtained on the preceding voiced frame or 20 ms whichever is the larger.

Once the width of the analysis window for a particular frame has been specified, the pitch-adaptive Hamming window, $w(n)$ is computed and normalized according to

$$\sum_{n=-N/2}^{N/2} w(n) = 1 \tag{2.4}$$

so that the periodogram peak will yield the amplitude of an underlying sine wave. Plotted in fig. 1 is a typical periodogram for voiced speech along with the amplitudes and frequencies that are estimated using the above procedure. The sine-wave phases are computed from the real and imaginary components of the STFT evaluated at sine-wave frequencies.

It should be noted that the placement of the analysis window $w(n)$ relative to the time origin is important for computing the phases. Typically in frame-sequential

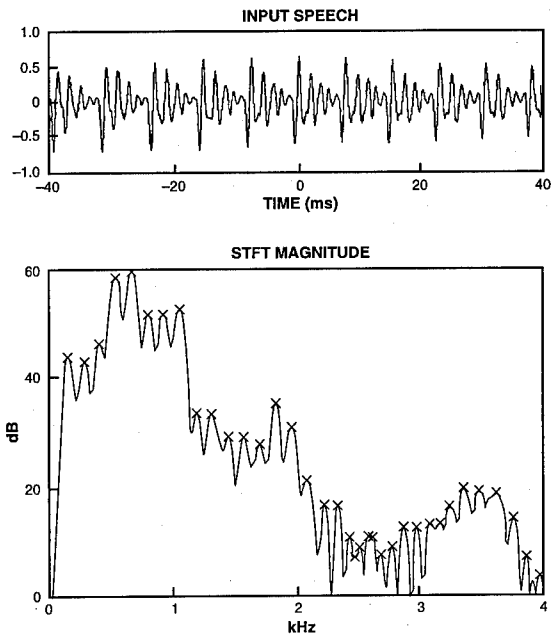


Figure 1. Typical periodogram for a frame of voiced speech and the amplitude and frequency estimates of the underlying sine waves.

processing the window, $w(n)$, lies in the interval $0 \leq n \leq N$, and is symmetric about $N/2$, a placement which gives the measured phase a linear term equal to $-\omega N/2$. Since N is on the order of 100–400 discrete time samples, any error in the estimated frequencies results in a large random phase error and consequent hoarseness in the reconstruction. An error of one DFT sample, for example, results in a $\frac{2\pi}{M}\frac{N}{2}$ phase error (where M is the DFT length) which could be on the order of π. To improve the robustness of the phase estimate the center of the Hamming window is placed at the origin defined as the center of the current analysis frame, corresponding to $n = 0$; hence the window takes on values over the interval $-N/2 \leq n \leq N/2$.

The approximations leading to the above periodogram estimator were based on the voiced speech assumption; nowhere have the properties of unvoiced speech been taken into account. To do this in a way that results in uncorrelated amplitude samples requires use of the Karhunen-Loève expansion for noise-like signals [17]. Such an analysis shows that a sinusoidal representation is valid provided the frequencies are "close enough" that the ensemble power spectral density changes slowly over consecutive frequencies. If the window width is constrained to be at least 20 ms

wide then, "on the average," there will be a set of periodogram peaks that will be approximately 100 Hz apart, and this provides a sufficiently dense sampling to satisfy the Karhunen-Loève constraints. Plotted in fig. 2 is a typical periodogram for a frame of unvoiced speech along with the amplitudes and frequencies that are estimated using the above procedure.

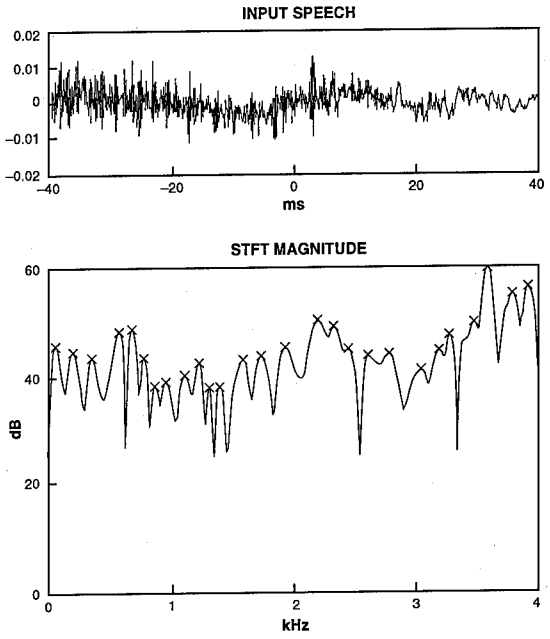


Figure 2. Typical periodogram for a frame of unvoiced speech and the amplitude and frequency estimates of the underlying sine waves.

The above analysis provides a heuristic justification for the representation of the speech waveform in terms of the amplitudes, frequencies, and phases of a set of sine waves that applies to one analysis frame. As speech evolves from frame to frame, different sets of these parameters will be obtained. The next problem to address then is the association of amplitudes, frequencies, and phases measured on one frame with those that are obtained on a successive frame in order to define sets of sine waves that will be continuously evolving in time.

2.3. Overlap-add sine-wave synthesis

If the amplitudes, frequencies, and phases that are estimated for the *kth* frame are denoted by $(A_\ell^k, \omega_\ell^k, \phi_\ell^k)$ then the synthetic speech for that frame can be computed

using

$$\hat{s}^k(n) = \sum_{\ell=1}^{L^k} A_\ell^k cos[n\omega_\ell^k + \theta_\ell^k] \tag{2.5}$$

Since the sine-wave parameters will be time-varying, discontinuities at the frame boundaries will be introduced unless provision is made for smoothly interpolating the parameters from one frame to those on another frame. Rather elegant methods have been developed for performing this task that involve establishing sine-wave tracks by matching the sine-wave frequencies from frame to frame and then using a linear function to interpolate the amplitudes and a cubic phase function to interpolate the frequencies and phase [3, 8]. Although solving the problem of synthesizing speech waveforms from a down-sampled set of sine-wave parameters, the method is not without significant computational expense particularly in the time required to execute the matching algorithm. Some systems attempt to avoid this problem using harmonic matching [8, 10, 12], but frequency chirps can be introduced unless the system is designed to insure that the pitch does not change significantly between analysis frames. This requires that the matching not be used if the pitch changes too much ($\approx 10\%$) [51] or that the frame size be kept sufficiently small ($\approx$ 5 ms) [12]. While the more general approach to sine-wave synthesis is needed for some applications such as signal separation for speech [12] and for music [23], the overlap-add interpolator [20, 21] eliminates the need to establish sine-wave tracks, is much simpler to implement and is perfectly satisfactory for the speech coding task provided the synthesis frame is made short enough that the sine-wave parameters satisfy the stationarity assumption.

In this case the synthetic speech waveform is obtained by applying eq. (2.5) to the sine-wave data on frames $k-1$ and k to generate the waveforms $\hat{s}^{k-1}(n)$ and $\hat{s}^k(n)$ respectively and these are each appropriately weighted, overlapped and added. Computationally this is equivalent to

$$\hat{s}(n) = w_s(n)\hat{s}^{k-1}(n) + w_s(n-T)\hat{s}^k(n-T) \tag{2.6}$$

where $w_s(n)$ is the overlap-and-add synthesis window that is designed such that

$$w_s(n) + w_s(n-T) = 1 \tag{2.7}$$

Triangular, Hanning and trapezoidal windows have typically been used for the sine-wave interpolation process.

2.4. Experimental results

In order to determine the effectiveness of the proposed sine-wave model a non-real-time floating-point simulation was developed using the analysis/synthesis system shown in fig. 3.

The speech processed in the simulation was low-pass-filtered at 4 kHz, digitized at 8 kHz, and analyzed at 10 ms frame intervals. A 512-point FFT using a pitch-adaptive Hamming window, having a width which was two and one-half times the average pitch gave accurate peak estimation for both voiced and unvoiced speech provided the window was at least 20 ms wide. The maximum number of peaks that was used in synthesis was set to a fixed number and, if excessive peaks were obtained, only the peaks corresponding to the first 100 frequencies were used. As shown in fig. 3 the triangular window was used in the overlap-and-add synthesis procedure.

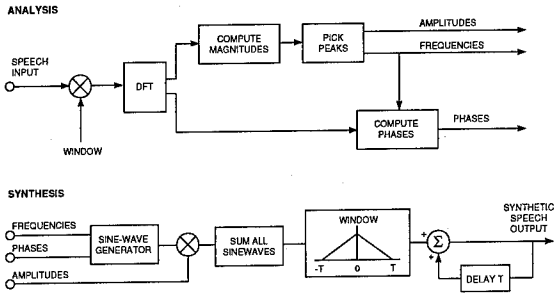


Figure 3. Block diagram of sine-wave analysis/synthesis system.

A large speech data base has been processed with this system, and it has been found that the synthetic speech was perceived to be essentially indistinguishable from the original. Visual examination of many of the reconstructed passages shows that the waveform structure is essentially preserved. This suggests that the quasi-stationarity conditions are satisfactorily met and that the use of the parametric model based on the amplitudes, frequencies, and phases of a set of sine-wave components appears to be justifiable for both voiced and unvoiced speech. Furthermore, when the overlap-and-add method was used in place of the matching and cubic-phase interpolation algorithm, there was no loss in performance provided the synthesis frame size was no greater than $\approx$ 12.5 ms. For a synthesis frame size T, the overlap-and-add method implicitly assumes that the sine-wave parameters are stationary over a window that is $2T$ in length. For T corresponding to a frame size greater than the 12.5 ms frame, stationarity would have to hold longer than 25 ms and this is clearly beyond the limits of the speech production mechanism. Experiments were performed using different windows satisfying eq. (2.7), such as the triangular and Hanning windows, but no discernible difference was perceived. In another experiment the inverse FFT was used to compute eq. (2.5) and, provided the FFT length was at least 512 points, for 4 kHz bandwidth speech no loss in quality was detected.

Although the sinusoidal model was originally designed for a single speaker, the general aharmonic model works equally well for reconstructing multi-speaker waveforms, music, speech in a musical background, and marine biologic signals such as whale sounds. Furthermore, it was found that the reconstruction did not break down in the presence of noise. The synthesized noisy speech is essentially perceptually indistinguishable from the original with no modification of the noise characteristics providing experimental justification for the validity of the Karhunen-Loève representation for noise-like signals.

Sine-wave analysis/synthesis has had a number of successful applications including time-scale and pitch-scale modification [22], peak-to-rms reduction [19], and two-talker separation [18]. It has been used for computer music synthesis [23] and for the analysis of vibrato [24]. More recently the decomposition of speech into sine-wave tracks has been used by Kleijn and Haagen [12, 13] to provide a basis for separating speech into slowly- and rapidly-varying components.

However the basic model that represents speech in terms of sets of sine-wave amplitudes, frequencies and phases turns out not to be amenable to low rate speech coding because there are simply too many parameters to be coded. In order to compress the data rate, therefore, the class of input signals must be restricted to speech so that more structured models for the sine-wave parameters can be used. In the next section a harmonic model for the sine-wave frequencies will be derived that will lead to a sine-wave based pitch extraction algorithm, which when used with the harmonic samples of the amplitude and phase of the STFT leads to high-quality synthetic speech. Then in section 4 a minimum-phase harmonic speech model will be developed that avoids the problem of working with the measured amplitudes and phases. The issues that arise in quantizing the parameters of the minimum-phase model to achieve performance at 4.8 kb/s and 2.4 kb/s will then be discussed in section 5.

3. A model for the sine-wave frequencies

The first step in the development of a low-rate sine-wave speech coder is to develop a model for the sine-wave frequencies. The most efficient model is based on the assumption that the sine-waves are harmonically related and then the problem is to estimate the frequency of the fundamental such that the harmonic set of sine waves is a "best fit" to the measured set of sine waves [25]. During voiced speech the estimated frequency can be interpreted as the speaker's pitch and the accuracy of the harmonic fit becomes an measure of degree to which the analyzed speech segment is voiced. During unvoiced speech the fundamental frequency has no physical meaning, but, with careful design of the estimation and synthesis procedures, it can lead to an effective sine-wave representation for speech in the unvoiced state. While there are now available time-domain pitch estimation algorithms that produce accurate and reliable pitch tracks that could have been used to determine the fundamental frequency of the harmonic sine-wave representation [26], the object here is to see if the frequency-domain approach could lead to new insights into the

pitch estimation and voicing detection problems.

3.1. Parameter estimation for the harmonic sine-wave model

As a first step in the analysis procedure, it is assumed that a frame of the input speech waveform has already been analyzed in terms of its sinusoidal components using the techniques described in section 2. The measured speech data, $s(n)$ can therefore be represented as

$$s(n) = \sum_{\ell=1}^{L} A_\ell exp[j(n\omega_\ell + \theta_\ell)] \tag{3.1}$$

where $\{A_\ell, \omega_\ell, \theta_\ell\}_{\ell=1}^{L}$ represent the amplitudes, frequencies, and phases of the L measured sine waves [1]. The goal is to try to represent this sinusoidal waveform by another waveform for which all of the frequencies are harmonic. This latter waveform can be modeled as

$$\hat{s}(n;\omega_0,\phi) = \sum_{k=1}^{K(\omega_0)} \bar{A}(k\omega_0) exp[j(nk\omega_0 + \phi_k)] \tag{3.2}$$

where $\omega_0 = 2\pi f_0/f_s$ is the normalized fundamental frequency, $K(\omega_0)$ is the number of harmonics in the speech bandwidth, $\bar{A}(\omega)$ is the vocal tract envelope, and $\phi = (\phi_1, \phi_2, \cdots, \phi_{K(\omega_0)})$ represents the phases of the harmonics. Henceforth, ω_0 will be referred to as the "pitch", although during unvoiced speech this terminology is not meaningful in the usual sense. It is desired to estimate the pitch frequency ω_0 and the phases $(\phi_1, \phi_2, \cdots, \phi_{K(\omega_0)})$ such that $\hat{s}(n)$ is as "close as possible" to $s(n)$ according to some meaningful criterion. A reasonable estimation criterion is to seek the minimum of the mean-squared-error (MSE),

$$\epsilon(\omega_0,\phi) = \frac{1}{N+1} \sum_{n=-N/2}^{N/2} |s(n) - \hat{s}(n;\omega_0,\phi)|^2 \tag{3.3}$$

over ω_0 and ϕ. The MSE in eq. (3.3) can be expanded as

$$\epsilon(\omega_0,\phi) = \frac{1}{N+1} \sum_{n=-N/2}^{N/2} \{|s(n)|^2 - 2Re[s(n)\hat{s}^*(n;\omega_0,\phi)] + |\hat{s}(n;\omega_0,\phi)|^2\} \tag{3.4}$$

[1] The analysis in this section is simplified by using the complex sine-wave representation

The first term of eq. (3.4) represents the power in the measured signal and is independent of the unknown parameters. It is denoted by

$$P_s = \frac{1}{N+1} \sum_{n=-N/2}^{N/2} |s(n)|^2 \tag{3.5}$$

Substituting eq. (3.2) in the second term of eq. (3.4) leads to the relation

$$\sum_{n=-N/2}^{N/2} s(n)\hat{s}^*(n;\omega_0,\phi) = \sum_{k=1}^{K(\omega_0)} \bar{A}(k\omega_0)exp(-j\phi_k) \sum_{n=-N/2}^{N/2} s(n)exp(-jnk\omega_0) \tag{3.6}$$

Finally, substituting eq. (3.2) in the third term of eq. (3.4) leads to the relation

$$\frac{1}{N+1} \sum_{n=-N/2}^{N/2} |\hat{s}(n;\omega_0,\phi)|^2 \simeq \sum_{k=1}^{K(\omega_0)} \bar{A}^2(k\omega_0) \tag{3.7}$$

where the approximation is valid provided the analysis window satisfies the condition $(N+1) \gg 2\pi/\omega_0$, which is more or less assured by making the analysis window two and one-half times the average pitch period. This condition assumes that the average pitch has already been computed, an issue that will be addressed later in the section. Letting

$$S(\omega) = \frac{1}{N+1} \sum_{n=-N/2}^{N/2} s(n)exp(-jn\omega) \tag{3.8}$$

denote the short-time Fourier transform (STFT) of the input speech signal, and using this in eq. (3.6), then the expression for the MSE in eq. (3.4) becomes

$$\epsilon(\omega_0,\phi) = P_s - 2Re\{\sum_{k=1}^{K(\omega_0)} \bar{A}(k\omega_0)exp(-j\phi_k)S(k\omega_0)\} + \sum_{k=1}^{K(\omega_0)} \bar{A}^2(k\omega_0) \tag{3.9}$$

Since the phase parameters only affect the second term in eq. (3.9), the MSE will be minimized by choosing

$$\hat{\phi}_k = arg[S(k\omega_0)] \tag{3.10}$$

and the resulting MSE will be given by

$$\epsilon(\omega_0) = P_s - 2\sum_{k=1}^{K(\omega_0)} \bar{A}(k\omega_0)|S(k\omega_0)| + \sum_{k=1}^{K(\omega_0)} \bar{A}^2(k\omega_0) \tag{3.11}$$

The unknown pitch affects only the second and third terms in eq. (3.11), and these can be combined by defining

$$\rho(\omega_0) = \sum_{k=1}^{K(\omega_0)} \bar{A}(k\omega_0)[|S(k\omega_0)| - \frac{1}{2}\bar{A}(k\omega_0)] \tag{3.12}$$

and the MSE can then be expressed as

$$\epsilon(\omega_0) = P_s - 2\rho(\omega_0) \tag{3.13}$$

Since the first term is a known constant, the minimum-mean-squared-error (MMSE) is obtained by maximizing $\rho(\omega_0)$ over ω_0.

It is useful to manipulate this metric further by making explicit use of the sinusoidal representation of the input speech waveform. Substituting the representation in eq. (3.1) first in eq. (3.5) the measured signal power becomes

$$P_s = \sum_{\ell=1}^{L} A_\ell^2 \tag{3.14}$$

and then in eq. (3.8) the STFT becomes

$$S(\omega) = \sum_{\ell=1}^{L} A_\ell exp(j\theta_\ell)\ sinc(\omega_\ell - \omega) \tag{3.15}$$

where

$$sinc(x) = \frac{1}{N+1}\sum_{n=-N/2}^{N/2} exp(jnx) = \frac{sin[(N+1)\frac{x}{2}]}{(N+1)sin(\frac{x}{2})} \tag{3.16}$$

Since the sine waves are well-resolved, the magnitude of the STFT can then be approximated by

$$|S(\omega)| \approx \sum_{\ell=1}^{L} A_\ell D(\omega_\ell - \omega) \tag{3.17}$$

where

$$D(x) = \begin{cases} sinc(x) & \text{if } |x| \le \frac{2\pi}{N+1} \\ 0 & otherwise \end{cases} \tag{3.18}$$

The optimization criterion then becomes

$$\rho(\omega_0) = \sum_{k=1}^{K(\omega_0)} \bar{A}(k\omega_0)[\sum_{\ell=1}^{L} A_\ell D(\omega_\ell - k\omega_0) - \frac{1}{2}\bar{A}(k\omega_0)] \tag{3.19}$$

To gain some insight into the meaning of this criterion, suppose that the input speech is periodic with pitch frequency ω^*. Then $\omega_\ell = \ell\omega^*$, $A_\ell = \bar{A}(\ell\omega^*)$ and

$$\rho(\omega^*) = \frac{1}{2}\sum_{k=1}^{K(\omega^*)} [\bar{A}(k\omega^*)]^2 \tag{3.20}$$

When ω_0 corresponds to submultiples of the pitch, the first term in eq. (3.19) remains unchanged, since $D(\omega_\ell - k\omega_0) = 0$ at the submultiples; but the second term, because it is an envelope and always non-zero, will increase at the submultiples of ω^*. As a consequence

$$\rho(\frac{\omega^*}{m}) < \rho(\omega^*) \quad m = 2, 3, \cdots \tag{3.21}$$

which shows that the above optimization criterion leads to unambiguous pitch estimates. This is possibly its most significant attribute, as it has been found through extensive experimentation that the usual problems with pitch period doubling do not occur with this metric. However, the frequency domain implementation can lead to additional processing advantages, the first of which is pitch-adaptive resolution.

3.2. Pitch-adaptive resolution

In the above formulation it was implied that the analysis window was fixed at $N+1$ samples. This would mean that the main lobe of the sinc-function, which measures the distance of the measured sine-wave frequencies from the harmonic candidates (i.e., $sinc(\omega_\ell - k\omega_0) \sim (\omega_\ell - k\omega_0)^2$ for $|\omega_\ell - k\omega_0|$ small) would be fixed for all pitch candidates. This is contrary to the fact that the ear is perceptually more tolerant to larger errors in the pitch at high pitch frequencies than at lower pitch frequencies. Moreover, the sinc-function distance measure of the error is meaningful only over each harmonic lobe. These effects can be accounted for by defining the distance function $D(x)$ at the kth harmonic lobe to be

$$D(\omega - k\omega_0) = \frac{sin[2\pi(\frac{\omega - k\omega_0}{\omega_0})]}{2\pi(\frac{\omega - k\omega_0}{\omega_0})} \qquad \text{for all } |\omega - k\omega_0| \leq \frac{\omega_0}{2} \tag{3.22}$$

and to be zero elsewhere. In this way the resolution becomes very sharp at low pitch values, and in contrast, becomes quite broad at high values of the pitch. It is this expression which is used in eq. (3.19) to compute the first revision to the optimality criterion.

3.3. Enhanced discrimination

The MSE criterion is closely related to the design of a Gaussian classifier for which the classes, the pitch candidates, are assumed to be independent. It is desirable that the classification algorithm not only detect the correct class with high probability, but also suppress the likelihood that any other class might be detected. This feature, which in a neural net classifier is known as negative reinforcement [27] can be incorporated into the MSE pitch estimation algorithm by noting that if ω_0 were the true pitch, then there would be at most one measured sine wave in each harmonic lobe tuned to ω_0. Therefore, if there are more, then only the one that contributes most to the MSE should be computed. Since the lobes are determined by the pitch-adaptive sinc-function in eq. (3.22) and, since each lobe spans one harmonic interval defined by the set

$$L(k\omega_0) = \{\omega : k\omega_0 - \frac{\omega_0}{2} \leq \omega < k\omega_0 + \frac{\omega_0}{2}\} \tag{3.23}$$

then discrimination will be enhanced by allowing only the largest weighted sine wave for each harmonic lobe. The second revision to the pitch optimality criterion is

$$\rho(\omega_0) = \sum_{k=1}^{K(\omega_0)} \bar{A}(k\omega_0)\{ \max_{\omega_\ell \epsilon L(k\omega_0)} [\, A_\ell D(\omega_\ell - k\omega_0)\,] - \frac{1}{2}\bar{A}(k\omega_0)\} \tag{3.24}$$

In addition to providing greater robustness against additive noise (since the small peaks due to noise are ignored), the enhanced MSE criterion insures that speech of low pitch will less likely be estimated as a high pitch. Moreover, if the above implementation is thought of as a form of small-signal-suppression and, if the harmonic lobe structure is thought of as an auditory critical band filter, then it is possible to speculate that enhanced discrimination is not unlike the effect of auditory masking of small tones by nearby large tones

3.4. The formant interaction problem

One of the more important pitch estimation techniques in current use is based on the correlation function. In some respects it is the time domain duality to the correlation implicit in the first term in eq. (3.19). One problem with the time-domain correlation technique is the result of the interaction between the pitch and the first formant. If the formant bandwidth is narrow relative to the harmonic spacing, the correlation function reflects the formant frequency rather than the underlying pitch. By inverse filtering the speech waveform and modifying the computation of the correlation function, the formant interaction problem no longer limits the performance of contemporary time-domain pitch estimators [26].

In the frequency domain the formant interaction problem arises as the sine-wave amplitude closest to the formant frequency tends to dominate the MSE criterion and lead to an ambiguous pitch estimate if that sine-wave is other than the first harmonic. This effect can be eliminated by reducing the dynamic range of all of the sine-wave amplitudes and, in turn, the amplitude envelope. One way to do this is to replace the measured sine-wave amplitudes by

$$A_\ell = \left(\frac{A_\ell}{A_{max}}\right)^\gamma \qquad 0 < \gamma \leq 1 \tag{3.25}$$

where $A_{max} = max\{A_\ell\}_{\ell=1}^{L}$. Since the MSE criterion leads to maximal robustness against additive white Gaussian noise, it was desirable to keep γ as close to unity as possible, introducing just enough amplitude compression to eliminate the formant interaction problem. Too much compression causes the low level peaks due to noise to distort the MSE criterion. Ultimately, the compression factor was chosen experimentally to be $\gamma = .5$.

3.5. Sine-wave amplitude envelope estimation

It has been shown that if the envelope of the sine-wave amplitudes is known, then the MSE criterion can lead to unambiguous estimates of the pitch. While a number of methods might be used for estimating the envelope using linear prediction or cepstral estimation techniques, for example, it was desirable to use a method that led to an envelope that passed through the measured sine-wave amplitudes. Such a technique has already been developed in the *spectral envelope estimation vocoder* (SEEVOC) [28].

The SEEVOC algorithm depends on having an estimate of the average pitch, denoted here by $\bar{\omega}_0$. The first step is to search for the largest sine-wave amplitude in the interval $[\frac{\bar{\omega}_0}{2}, \frac{3\bar{\omega}_0}{2}]$. Having found the amplitude and frequency of that peak, labeled (A_1, ω_1), then the interval $[\omega_1 + \frac{\bar{\omega}_0}{2}, \omega_1 + \frac{3\bar{\omega}_0}{2}]$ is searched for its largest peak, labeled (A_2, ω_2). The process is continued by searching the intervals $[\omega_{\ell-1} + \frac{\bar{\omega}_0}{2}, \omega_{\ell-1} + \frac{3\bar{\omega}_0}{2}]$ for the largest peaks, (A_ℓ, ω_ℓ) until the edge of the speech bandwidth is reached. If no peak is found in a search bin, then the value of the short-time Fourier transform (STFT) magnitude at the bin center is used and becomes the point from which the search procedure is continued. The principle advantage of this method is the fact that any low-level peaks within a harmonic interval will be masked by the largest peak, presumably a peak that is close to an underlying harmonic. Moreover, the procedure is not dependent on the peaks being harmonic, nor on the exact value of the average pitch since the procedure resets itself after each peak has been found. The SEEVOC envelope is then obtained by applying piecewise constant interpolation between the sine-wave amplitudes and frequencies that were obtained using the SEEVOC peak-picking routine. It is this envelope that is used for $\bar{A}(\omega)$ in the evaluation of the pitch estimation criterion in eq. (3.24).

3.6. Two-pass pitch estimation

The MSE pitch extractor is predicated on the assumption that the input speech waveform has been represented in terms of the sinusoidal model. This implicitly assumes that the analysis has been performed using a Hamming window approximately two and one-half times the average pitch. Moreover, the SEEVOC technique also assumes that an estimate of the average pitch is available. It seems, therefore, that the pitch has to be known in order to estimate the average pitch, in order to estimate the pitch. This circular dilemma can be broken by using the sine-wave based pitch estimator on a fixed window of width two and one-half times the largest pitch period. The mean-squared-error measure of the quality of the harmonic fit can be used to control the up-date of the average pitch which in turn is used in computing the SEEVOC envelope. The estimated pitch is then used to set the pitch-adaptive window which is necessary to get the best estimate of the sine-wave amplitudes prior to coding. Since this requires that the entire STFT analysis be repeated, the pitch extraction algorithm could also be repeated using the sine-wave parameters obtained using the pitch-adaptive window. However this proves to be computational expensive for single-chip DSP applications. The complexity can be reduced significantly by restricting the search range of the second search to a small neighbourhood about the pitch estimated on the fixed wide window. Little, if any, performance loss has been observed using the pitch refinement technique. Moreover, since the purpose of the first search is to estimate only the pitch, the STFT analysis needs to be performed over only the $1000Hz$ baseband since it is in this region that the harmonic structure is most reliable during voiced speech.

In order to provide the SEEVOC algorithm with an average pitch, it is necessary to determine when the pitch represents an essentially voiced frame. This will be the subject discussed in the next section.

3.7. Voicing detection

In the context of the sinusoidal model the degree to which a given frame of speech is voiced is determined by the degree to which the harmonic model fits the original sine-wave data. The accuracy of the harmonic fit can be related, in turn, to the signal-to-noise ratio (SNR) defined by

$$SNR = \frac{\sum_n |s(n)|^2}{\sum_n |s(n) - \acute{s}(n;\hat{\omega}_0)|^2} \tag{3.26}$$

where $\hat{\omega}_0$ is the pitch estimated using the procedures described in the previous sections. From eqs. (3.3) and (3.13) it follows that

$$SNR = \frac{P_s}{P_s - 2\rho(\hat{\omega}_0)} \tag{3.27}$$

where now the input power, P_s, is computed for the compressed sine-wave amplitudes that were defined in eq. (3.25). If the SNR is large, then the MSE is small and the harmonic fit is very good, which indicates that the input speech is most likely voiced. For small SNR, on the other hand, the MSE is large and the harmonic fit is quite poor which indicates that the input speech is more likely to be unvoiced. Therefore, the degree of voicing is functionally dependent on the SNR. Although the determination of the exact functional form is difficult to determine, one that has proven useful in several speech applications is the following:

$$P_v(SNR) = \begin{cases} 1 & SNR > 13 \text{ dB} \\ \frac{1}{9}(SNR - 4) & 4 \text{ dB} \leq SNR \leq 13 \text{ dB} \\ 0 & SNR < 4 \text{ dB} \end{cases} \tag{3.28}$$

where, P_v, the voicing level, represents the likelihood that the speech is voiced. The average pitch can be computed by using a simple first order filter updated whenever the voicing level is above some reasonable threshold ($\approx .8$). Pitch continuity constraints can also be added, allowing for a relaxation of the voicing threshold.

3.8. Experimental results

In one implementation of the MSE pitch extractor the speech was sampled at 8 kHz and Fourier analyzed using a 512-point FFT. The sine-wave amplitudes and frequencies were determined over a 1000 Hz bandwidth. The locations of the frequencies were refined using quadratic interpolation. In fig. 4(b), the measured amplitudes and frequencies are shown along with the piecewise-constant SEEVOC envelope. Square-root compression has been applied to the amplitude data. Figure 4(c) is a plot of the first term in eq. (3.19) over a pitch range from 38 Hz to 400 Hz and the inherent ambiguity of the correlator is apparent. It should be noted that "most of the time" the peak at the correct pitch has the largest value, but during steady vowels the ambiguous behavior illustrated in the figure commonly occurs. Figure 4(d) is a plot of the overall MSE criterion and the manner in which the ambiguities are eliminated is clearly demonstrated. Figure 5 illustrates typical results for a segment of unvoiced speech.

3.9. Harmonic sine-wave model

Validating the performance of a pitch extractor can be a time-consuming and laborious procedure since it requires a comparison with hand-labeled data. The approach used in the present study was to reconstruct the speech using the harmonic sine-wave model and to listen for pitch errors. The procedure is not quite so straightforward, however, since during unvoiced speech meaningless pitch estimates are made which can lead to perceptual artifacts whenever the pitch estimate is greater than about 150 Hz. This is due to the fact that in these cases there are too few sine waves to adequately synthesize a noiselike waveform. This problem has been eliminated

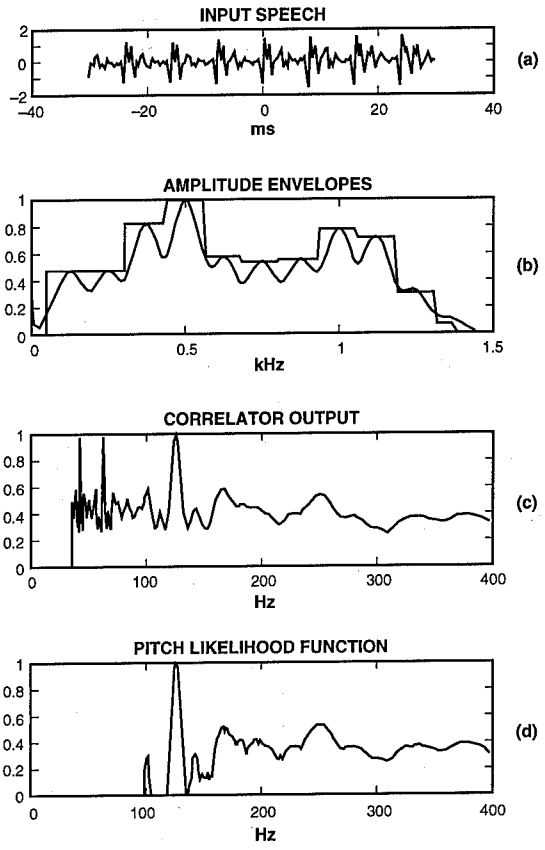


Figure 4. Typical pitch estimator results for voiced speech.

by defaulting to a fixed low pitch ($\approx$ 100 Hz) during unvoiced speech whenever the pitch exceeds 100 Hz. The exact procedure for doing this is to first define a voicing-dependent cutoff frequency, ω_c, as

$$\omega_c(P_v) = \pi P_v \tag{3.29}$$

which is constrained to be no smaller than 2π (1500 Hz/f_s). If the actual pitch estimate is ω_0, then the sine-wave frequencies used in the reconstruction are

$$\omega_k = \begin{cases} k\omega_0 & \text{for } k\omega_0 \leq \omega_c(P_v) \\ k^*\omega_0 + (k - k^*)\omega_u & \text{for } k\omega_0 > \omega_c(P_v) \end{cases} \tag{3.30}$$

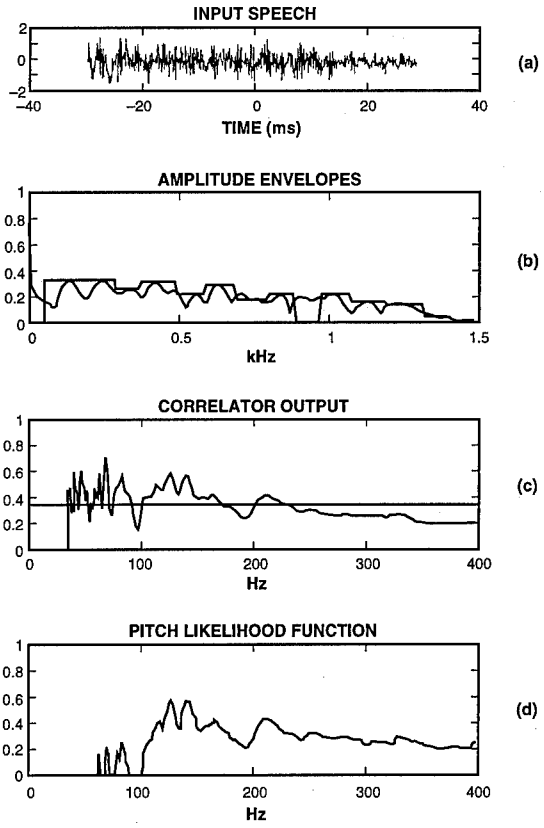


Figure 5. Typical pitch estimator results for unvoiced speech.

where k^* is the largest value of k for which $k^*\omega_0 \leq \omega_c(P_v)$, and where ω_u, the unvoiced pitch corresponds to 100 Hz (i.e., $\omega_u = 2\pi(100/f_s)$). Note that if $\omega_0 < \omega_u$, then $\omega_k = k\omega_0$ for all k. The harmonic reconstruction then becomes

$$\hat{s}(n;\omega_0) = \sum_{k=1}^{K} \bar{A}(\omega_k) exp[j(n\omega_k + \phi_k)] \tag{3.31}$$

where ϕ_k is the phase of the STFT at frequency ω_k. Strictly speaking, this procedure is harmonic only during strongly-voiced speech since if the speech is a voiced/unvoiced mixture the frequencies above the cutoff, although equally spaced by ω_u, are aharmonic, since they are themselves not multiples of the fundamental pitch.

The synthetic speech produced by this model is of very high quality, almost perceptually equivalent to the original, provided the frame-rate is less that ≈ 12.5 ms. Not only does this validate the performance of the MSE pitch extractor, but it also shows that if the amplitudes and phases of the harmonic representation could be efficiently coded, then only the pitch and voicing would be needed to code the information in the sine-wave frequencies. Methods have been developed for time-differentially encoding the sine-wave phases, but the resulting coder must operate at around 13 kb/s. In order to achieve low data rates, therefore, models have to be developed for the sine-wave phases. This will be one of the topics discussed in the next section.

4. Minimum phase harmonic sine-wave speech model

In the previous section it was shown that that synthetic speech of high quality could be synthesized using a harmonic set of sine waves provided the amplitudes and phases were the samples of the magnitude and phase of the short-time Fourier transform at each of the sine-wave harmonics. Even though the harmonic model eliminated the need to code the sine-wave frequencies, the amplitudes and phases would have to be quantized, and, in general, there remain too many parameters to encode and achieve operation at 4800 b/s and less. Therefore more of the properties of the speech production mechanism need to be exploited in order to further reduce the size of the parameter set. In this section models for the glottal excitation and vocal tract transfer function will be used to obtain a reduced parameter set for coding.

4.1. Voiced speech sine-wave model

During strongly voiced speech the production of speech begins with a sequence of excitation pitch pulses that represent the closure of the glottis at a rate given by the pitch frequency ω_0. Such a sequence can be written in terms of a sum of sine waves as

$$\hat{e}(n) = \sum_{\ell=1}^{L} exp[j(n - n_0)\omega_\ell] \tag{4.1}$$

where n_0 corresponds to the time of occurrence of the pitch pulse nearest the center of the current analysis frame. The occurrence of this temporal event, called the onset time, ensures that the underlying excitation sine waves will be in phase at the time of the occurrence of the glottal pulse. It is noted that although the glottis may close periodically, the measured sine waves may not be perfectly harmonic, hence the frequencies ω_ℓ may not in general be harmonically related to the pitch frequency.

The next operation in the speech production model shows that the amplitude and phase of the excitation sine waves are altered by the glottal pulse and vocal tract filters. Letting $H_s(\omega) = |H_s(\omega)|exp[j\Phi_s(\omega)]$ denote the composite transfer function for these filters, called the system function, then the speech signal at its output due to the excitation pulse train at its input can be written as

$$\hat{s}(n) = \sum_{\ell=1}^{L} |H_s(\omega_\ell)|exp[j(n-n_0)\omega_\ell + \Phi_s(\omega_\ell)] \tag{4.2}$$

Using the same measurements of the sine-wave parameters that were provided to the pitch estimation algorithm, the current frame of speech that is being analysed can be represented by the model

$$s(n) = \sum_{\ell=1}^{L} A_\ell exp[j(n\omega_\ell + \theta_\ell)] \tag{4.3}$$

The sine-wave amplitudes and phases correpsonding to the values that would have been produced by the above glottal and vocal tract models, can then be identified as:

$$\begin{aligned} A_\ell &= |H_s(\omega_\ell)| \\ \theta_\ell &= -n_0\omega_\ell + \Phi_s(\omega_\ell) \end{aligned} \tag{4.4}$$

This identifies the sine-wave amplitudes as samples of the magnitude of the vocal tract envelope, hence it should be possible to apply a suitable interpolation function so that the envelope can be estimated. One interpolation function would be the bandpass interpolator, but this is computationally complex and would make real-time implementation of the resulting algorithm difficult. However, it has been shown that good approximations to the ideal bandpass interpolator are available using cubic spline processing techniques [29], and these methods have been found to work well in the context of the sine-wave system provided the cubic spline envelope is fitted to the *logarithm* of the SEEVOC peaks. In the present context the SEEVOC algorithm [28] uses the current measurement of the pitch, denoted here by ω_0, to search for the largest sine-wave amplitude in the interval $[\frac{\omega_0}{2}, \frac{3\omega_0}{2}]$. Having found the amplitude and frequency of that peak, labeled (A_1, ω_1), then the interval $[\omega_1 + \frac{\omega_0}{2}, \omega_1 + \frac{3\omega_0}{2}]$ is searched for its largest peak, labeled (A_2, ω_2). The process is continued by searching the intervals $[\omega_{\ell-1} + \frac{\omega_0}{2}, \omega_{\ell-1} + \frac{3\omega_0}{2}]$ for the largest peaks, (A_ℓ, ω_ℓ) until the edge of the speech bandwidth is reached. If no peak is found in a search bin, then the value of the short-time Fourier transform magnitude at the bin center is used and becomes the point from which the search procedure is continued.

The principle advantage of this method is the fact that any low-level peaks within a harmonic interval will be masked by the largest peak, presumably a peak that is close to an underlying harmonic. Moreover, the procedure is not dependent on the peaks being harmonic, nor on the exact value of the pitch since the procedure

resets itself after each peak has been found. The estimate of the log-magnitude of the vocal tract transfer function be taken as the cubic spline fit to the *logarithm* of the sine-wave amplitudes at the frequencies obtained using the SEEVOC peak-picking routine. An example of such a fit to typical speech data is shown in fig. 6

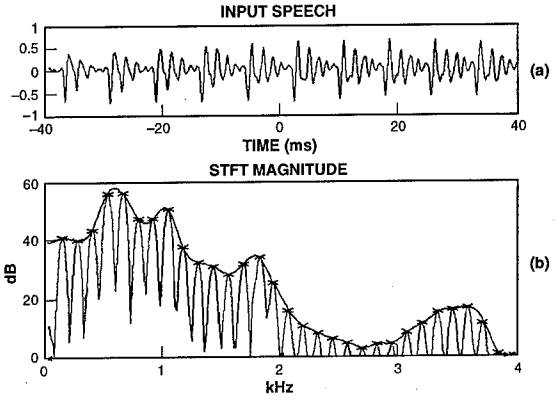


Figure 6. Cubic spline envelope fitted to the SEEVOC peaks.

If it is assumed that the vocal tract transfer function is minimum phase, which is the same phase model underlying the class of LPC and CELP based vocoders, then the magnitude and phase of $H_s(\omega)$ can be expressed in terms of a set of cepstral coefficients $\{c_m\}_{m=0}^{M}$ as [30]

$$\log|H_s(\omega)| = c_0 + 2\sum_{m=1}^{M} c_m \cos(m\omega) \tag{4.5}$$

$$\Phi_s(\omega) = -2\sum_{m=1}^{M} c_m \sin(m\omega)$$

$$c_m = \frac{1}{\pi}\int_0^{\pi} \log|H_s(\omega)| \cos(m\omega)d\omega \quad m = 0, 1, 2, ..., M$$

In practice, for a 4 kHz speech bandwidth, $M \geq 44$ is sufficient for achieving a good cepstral fit to the cubic spline envelope. If these results are now substituted into eq. (4.2) then, except for the onset time, all of the parameters of the harmonic minimum phase speech model are determined and depend only on the cubic spline fit to the sine-wave amplitudes.

Since the function of the onset time is to bring the sine waves into phase at times corresponding to the occurrence of a pitch pulse, then rather than attempt to estimate the onset time from the data, as was done in [15, 31, 32], it is possible to achieve the same perceptual effect simply by keeping track of successive onset times

generated by a succession of pitch periods that are available at the synthesizer. If the pitch period is stationary over the synthesis frame, and if n_0^{k-1} is the onset time for frame $k-1$ and P_0^{k-1} is the pitch period estimated for that frame, then a succession of synthetic onset times can be specified by

$$n_0^{k-1}(j) = n_0^{k-1} + jP_0^{k-1} \qquad j = 1, 2, \cdots, J \tag{4.6}$$

If $n_0^{k-1}(J)$ is the onset time closest to the mid-point between frame $k-1$ and frame k, then another sequence of onset times that better reflect the phase properties on frame k would be given by

$$n_0^k(i) = n_0^{k-1}(J) + iP_0^k \qquad i = 1, 2, \cdots, I \tag{4.7}$$

where P_0^k is the pitch period estimated for frame k. An example of a typical sequence of onset times is shown in fig. 7. Also shown is the fact that for high-pitched speakers there can be more than one onset time per analysis frame, and although any one of the onset times can be used, in the face of computational errors it is best to choose the onset time which is nearest the center of frame k.

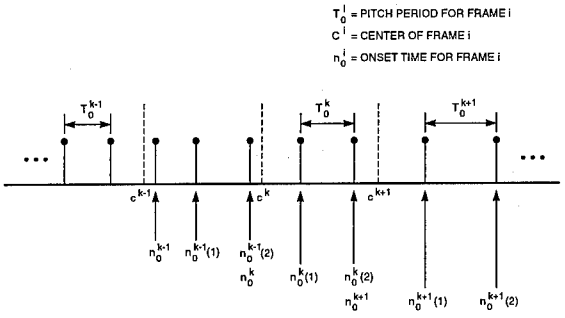


Figure 7. Sequence of onset times computed from successive pitch periods.

Another way to compute the onset times while accounting for the effects of time-varying pitch is to define the phase of the fundamental frequency to be the integral of the instantaneous frequency, viz.

$$\phi_0(t) = \phi_0[(k-1)T] + \int_{(k-1)T}^{t} \omega_0(\sigma)d\sigma \tag{4.8}$$

where $\omega_0(t)$ is the pitch frequency at time t. Since this phase will be monotonically increasing with t, a sequence of onset times can be found at the values of t for which $\phi_0(n_0) = 2\pi M$ for integer values of M. If ω_0^{k-1} and ω_0^k denote the estimated

pitch frequencies on frames $k-1$ and k respectively, then a reasonable model for the frequency variation in going from frame $k-1$ to frame k is

$$\omega_0(t) = \omega_0^{k-1} + \frac{\omega_0^k - \omega_0^{k-1}}{T} t \tag{4.9}$$

which can be used to compute the phase in eq. (4.8) and subsequently the onset time.

If all of the sine waves are harmonically related, then the phase of the $\ell' th$ sine wave is simply ℓ times the phase of the fundamental which means that the excitation sine waves will be in phase for every point in time. This leads to a phase model for which it is unnecessary to compute the onset time explicity simply by defining the phase offset for the $\ell' th$ sine wave as ℓ times the phase offset of the fundamental evaluated at the center of the current synthesis frame [3, 33]. Using eq. (4.9) in eq. (4.8), this phase will be given by

$$\phi_0(kT) = \phi_0[(k-1)T] + (\omega_0^{k-1} + \omega_0^k)(T/2) \tag{4.10}$$

and combining the excitation phase and system phase as in eq. (4.4), the voiced speech sine-wave phase for the $\ell' th$ harmonic becomes

$$\hat{\theta}(\ell\omega_0) = \ell\phi_0(kT) + \Phi_s(\ell\omega_0) \tag{4.11}$$

This shows that for voiced speech the sine-wave reconstruction depends only on the pitch, and, through the cubic spline envelope, the sine-wave amplitudes.

4.2. Unvoiced speech sine-wave model

If the above phase model is used in place of the measured sine-wave phases the synthetic speech is quite natural during voiced speech, but very buzzy during the unvoiced segments. On the other hand, if the phases are replaced by unformly distributed random variables on $[-\pi, \pi]$, then the speech is quite natural during unvoiced speech but sounds like whispered speech during the voiced segments. This suggests that the phase model in eq. (4.11) be generalized by adding a voicing-dependent component which would be zero for voiced speech and random on $[-\pi, \pi]$ for unvoiced speech. However it should be expected that as in standard low-rate "buzz/hiss" vocoders such a binary voiced/unvoiced phase model would render the sine-wave system overly dependent on the voicing decision causing similar artifacts to occur in the synthetic speech when this decision was made erroneously. The deleterious effects of the binary decision can be reduced significantly by using a more general mixed excitation model of the type proposed by Makhoul, et al., [34]. In their model, a voicing transition frequency was estimated below which voiced speech was synthesized and above which unvoiced speech was synthesized. Although their work was done in the context of a conventional LPC vocoder (periodic impulse

train for voiced speech, random noise for unvoiced speech), the concept is ideally suited to the sine-wave model, since below the voicing transition frequency, the sine-wave phase residuals are made zero, and above the transition they are made random [33, 35]. Letting ω_c denote the voicing-dependent cutoff frequency then the unvoiced phase component can be modelled by

$$\hat{\epsilon}(\omega) = \begin{cases} 0 & \text{if } \omega \leq \omega_c \\ U[-\pi, \pi] & \text{if } \omega > \omega_c \end{cases} \tag{4.12}$$

where $U[-\pi, \pi]$ denotes a uniformly distributed random variable on $[-\pi, \pi]$. If this is added to the voiced-speech phase model in eq. (4.11) the complete sine-wave phase for the $\ell' th$ harmonic becomes

$$\hat{\theta}(\ell\omega_0) = \ell\phi_0(kT) + \Phi_s(\ell\omega_0) + \hat{\epsilon}(\ell\omega_0) \tag{4.13}$$

To capture the random phase during mixed unvoiced speech the phase model requires the computation of the voicing-adaptive cutoff frequency ω_c in eq. (4.12). One approach might be to fit the above phase model to the measured sine-wave phases while trying to determine the cutoff frequency above which the residual phase became random. Attempts have been made to do this [31, 15] but the method is computational complex and very sensitive to the phase characteristics of the input audio processing. Based on the analysis in section 3 an alternate measure of voicing was obtained that measured the fit of the harmonic set of sine waves to the measured set of sine waves. Using this measure, which was called the voicing probability P_v, from eq. (3.28), the voicing-dependent cutoff frequency can then be estimated as $\omega_c(P_v) = \pi P_v$.

Therefore the minimum-phase harmonic sine-wave synthesis model generated using parameters estimated for the $k' th$ frame can be summarized as follows:

$$\begin{aligned} \hat{s}(n) &= \sum_{\ell=0}^{L} \hat{A}_\ell exp[j(n\ell\omega_0 + \hat{\theta}_\ell)] \\ \hat{A}_\ell &= |H_s(\ell\omega_0)| \\ \hat{\theta}_\ell &= \begin{cases} \ell\phi_0(kT) + \Phi_s(\ell\omega_0) & \text{if } \ell\omega_0 \leq \omega_c \\ \ell\phi_0(kT) + \Phi_s(\ell\omega_0) + U[-\pi, \pi] & \text{if } \ell\omega_0 > \omega_c \end{cases} \end{aligned} \tag{4.14}$$

where $|H_s(\omega)|$ is the magnitude of the vocal tract tract transfer function and is obtained by fitting a cubic spline envelope to the SEEVOC-edited sine-wave amplitudes, $\Phi_s(\omega)$ is the vocal tract phase determined from the vocal tract amplitude envelope through the minimum phase assumption and $\phi_0(kT)$ is the phase of the fundamental at the center of the $k' th$ synthesis frame determined by the phase-locking procedure that led to eq. (4.10). The sine-wave phases are made random above the voicing-adaptive cutoff frequency, $\omega_c = \pi P_v$, which is determined by the the voicing probability P_v that is a measure of how well the harmonic set of sine waves fits the measured set of sine waves and was determined as part of the

pitch estimation process in eq. (3.28). As in the basic sine-wave reconstruction system described in section 2, speech is synthesized over contiguous frames using the overlap-add algorithm with triangular weighting and this requires that the analysis and synthesis procedures be updated at least once per 12.5 ms frame.

4.3. Postfilter design

While the synthetic speech produced by this system was of reasonably good quality, a muffling effect could be detected particularly for certain low-pitched speakers. Such a quality loss has also been reported in code-excited LPC systems where it has been argued that the muffling is due to coder noise in the formant nulls. Techniques have been developed for filtering out this noise by passing the synthesized speech through a postfilter [36]. Since the synthetic speech produced by the minimum phase harmonic sine-wave system has not been quantized, the muffling cannot be attributed to quantization noise, but to the front-end analysis that led to the sine-wave amplitude representation in the first place. Instead of quantization noise filling in the formant nulls, it is speculated that the degradation occurs due to sidelobe leakage. Since the Hamming window has a 40 dB sidelobe level, sine waves near the formant peaks can easily dominate the formant null reducing the dynamic range of a formant peak-to-null to be no more than 40 dB. A variant of the CELP postfilter design technique has been developed for sine-wave systems [15, 37] that uses a frequency-domain design approach to deepen the formant nulls. That method will now be discussed.

Basically the postfilter is a normalized, compressed version of the spectrally flattened vocal tract envelope, which when applied to the vocal tract envelope results in formants having deeper nulls that in turn result in synthetic speech that is less muffled. If $T(\omega)$ measures the spectral tilt of the vocal tract envelope $H_s(\omega)$, then

$$F(\omega) = \frac{|H_s(\omega)|}{|T(\omega)|} \tag{4.15}$$

represents a spectrally flattened version of it. If $F(\omega)$ is normalized to have unity gain, denoted by $\bar{F}(\omega)$, then it can be compressed using a root-γ compression rule, which defines the postfilter as

$$P(\omega) = \left[\bar{F}(\omega)\right]^{\gamma} \qquad 0 \leq \gamma \leq 1. \tag{4.16}$$

A simple model for the spectral tilt is the first-order all-pole model

$$T(\omega) = \frac{\sigma}{1 - \rho \exp(-j\omega)} \tag{4.17}$$

where the prediction coefficient ρ can be determined by applying LPC analysis techniques to the synthetic speech waveform in eq. (4.14). It then follows that

$\rho = R_1/R_0$, where R_0 and R_1 are the energy and correlation coefficient of $\hat{s}(n)$. It is easy to show that

$$R_0 = \sum_{\ell=1}^{L} \hat{A}_\ell^2$$

$$R_1 = \sum_{\ell=1}^{L} \hat{A}_\ell^2 \cos(\ell\omega_0) \tag{4.18}$$

hence the spectrally flattened sine-wave amplitude for the $\ell' th$ harmonic, $F(\ell\omega_0)$, can be written as

$$F(\ell\omega_0) = \frac{\hat{A}_\ell}{|T(\ell\omega_0)|} = \hat{A}_\ell \left[\frac{(1+\rho^2) - 2\rho\cos(\ell\omega_0)}{\sigma^2}\right]^{\frac{1}{2}} \tag{4.19}$$

The gain, σ, is chosen so that the average power of the spectrally flattened sine-wave amplitudes is unity, which requires that

$$\frac{1}{L}\sum_{\ell=1}^{L} F(\ell\omega_0)^2 = 1 = \frac{1}{L\sigma^2}[(1+\rho^2)\sum_{\ell=1}^{L}\hat{A}_\ell^2 - 2\rho\sum_{\ell=1}^{L}\hat{A}_\ell^2\cos(\ell\omega_0)] \tag{4.20}$$

Using the definitions in eq. (4.18) and solving for σ leads to

$$\sigma^2 = \frac{1}{L}[(1+\rho^2)R_0 - 2\rho R_1] \tag{4.21}$$

After some algebra it is easy to show that the spectrally-flattened sine-wave amplitudes can be written as

$$\bar{F}(\ell\omega_0) = \hat{A}_\ell \left[\frac{L[R_0^2 + R_1^2 - 2R_0R_1\cos(\ell\omega_0)]}{R_0(R_0^2 - R_1^2)}\right]^{\frac{1}{2}} \tag{4.22}$$

The post-filter weight at the $\ell' th$ harmonic is the unity gain spectrally-flattened sine-wave amplitude raised to the γ power. Letting these weights be denoted by W_ℓ it follows that

$$W_\ell = [\bar{F}(\ell\omega_0)]^\gamma = \hat{A}_\ell^\gamma \left[\frac{L[R_0^2 - 2R_1R_0\cos(\ell\omega_0) + R_1^2]}{R_0(R_0^2 - R_1^2)}\right]^{\frac{\gamma}{2}} \tag{4.23}$$

where it is noted that L, the number of harmonics is given by the integer value that is less than or equal to π/ω_0. In order to insure that excessive spectral shaping is not applied to any one sine-wave, clipping rules are introduced such that the final post-filter values at each harmonic are

$$P(\ell\omega_0) = \begin{cases} 1.2 & \text{if } W_\ell > 1.2 \\ 0.5 & \text{if } W_\ell < 0.5 \\ W_\ell & \textit{otherwise} \end{cases} \tag{4.24}$$

and the post-filtered sine-wave amplitude at the $\ell'th$ harmonic is then

$$\tilde{A}_\ell = P(\ell\omega_0)\hat{A}_\ell \tag{4.25}$$

In order maintain the correct energy level in the synthetic speech the post-filtered amplitudes are replaced by $\alpha\tilde{A}_\ell$ where the scale factor α is chosen such that the energy in the post-filtered waveform is the same as that before postfiltering. This requires that

$$\alpha = \left[\frac{R_0}{\sum_{\ell=0}^{L}\tilde{A}_\ell^2}\right]^{\frac{1}{2}} \tag{4.26}$$

It has been observed that sometimes the spectral nulls are reduced at the expense of some distortion in the amplitudes at the first formant peak, a problem that would not have arisen had the spectral tilt provided a better fit to the formant peaks. An alternative to using the first order all-pole model is to fit the cepstral tilt

$$\log|T(\omega)| = c_0 + 2c_1\cos(\omega) \tag{4.27}$$

to the log-sine-wave amplitudes and then repeat the steps described above. While this method results in better spectral flattening with less attenuation in the high-frequency end producing synthetic speech that sounds closer to the original, it is with the introduction of high-frequency artifacts, which overall, are perceptually less pleasant. It is clear that more work needs to be done to develop a better understanding of the spectral flattening and the post-filtering procedures, but in the interim, the postfilter based on the allpole model appears to be the preferred approach.

4.4. Experimental results

When the voicing-dependent synthetic phase model was used to replace the measured phases in the harmonic sine-wave synthesizer, not only was the speech of very high quality, but the "synthetic speaker" sounded like the input speaker, i.e. the speaker identification properties were preserved. It was particularly notable that the synthetic speech did not have the "reverberant" quality reported in other implementations of the sine-wave system [56], an effect which arises when the component sine waves are not forced to be phase locked. Moreover, the effect of the postfilter was to make the synthetic speech sound more crisp, removing the "muffled" quality that seems to be inherent in the entire class of low-rate speech coders.

Comparisons were made using the minimum-phase versus a zero-phase model that is used in some implementations of the sine-wave system. In almost all cases, except perhaps for very high-pitch speakers, the minimum phase model adds the desired dispersive characteristics to the sine-wave phases making the synthetic speech sound more natural. This may be particularly important during mixed voiced-unvoiced

speech segments, since the inherent randomness of the sine-wave amplitudes is transferred to the system phase through the minimum phase assumption. This adds randomness to the appropriate sine-wave phases and, hence, more naturalness to the synthetic speech. In contrast, the zero-phase system was somewhat more muffled and less natural sounding.

Since the parameters of the minimum-phase harmonic speech model are the pitch frequency, the voicing probability and the cubic spline envelope, and since relatively few bits would be required to encode the pitch and voicing, the ability to operate such a synthesizer at low data rates depends on the number of bits required to encode the cubic spline envelope. The development of such an encoding strategy will be the topic of the next section.

5. Sine-wave amplitude coding using an all-pole model

So far it has been shown that the post-filtered minimum-phase harmonic model can produce synthetic speech having a very high level of quality that would certainly be acceptable for operation at low data rates. Since the pitch can be coded using $\approx$ 6-7 bits and the voicing probability using $\approx$ 2-3 bits, then low-rate operation appears to be achievable provided the spline envelope can be coded efficiently. Paul [28] addressed a similar problem in the development of the *spectral envelope estimation vocoder* (SEEVOC) by treating the spline envelope as a waveform which, by down-sampling and low-pass filtering, could be represented with a minimal set of samples which were then encoded differentially in frequency. The approach taken in this section is to fit a parametric model to the spline envelope and then code the parameters of the model. Previous work using this approach in the context of the sine-wave system has been explored extensively in the context of the cepstral model [15, 37]. The advantage of the cepstral representation is that it assumes no constraining model shape, except that the spline envelope represent a vocal tract transfer function that is minimum phase. Using time-differential and frequency-differential encoding methods it was possible to achieve reasonably good performance at 4800 b/s, but at 2400 b/s the synthetic speech was marginally acceptable. In this section the vocal tract transfer function will be further constrained to be minimum phase *and* all-pole. Building on the ideas originally suggested in [37–39], it will be shown that the more constrained model leads to quantization rules that are more bit-rate efficient than those obtained using the cepstral modeling methods.

5.1. The all-pole model

Although the all-pole model is used implicitly in the class of LPC and CELP based coders, the parameters of the model depend on a set of correlation coefficients that are computed directly from the time-domain waveform. The problem with this approach is that the speech waveform is periodic which renders the correlation function periodic. In order to insure that the corresponding spectral envelope does not

resolve the underlying sine-wave frequencies, the number of correlation coefficients must be restricted to one-half the smallest pitch period. At 8000 Hz sampling, a 400 Hz pitch has a period of 2.5 ms or 20 samples, thereby limiting the all-pole fit to a tenth order model. Such a low model order can cause the envelope to smooth the detail in the baseband, distort the speaker identifiability, and lead to a phase function that is so structured that, in contrast to the unconstrained minimum-phase system, results in synthetic speech that sounds mechanical and buzzy.

An alternative to the time-domain approach is to use the basic sine-wave system to estimate the sine-wave amplitudes and then fit the all-pole model to those amplitudes. This approach has been studied in the case of voiced speech by McAulay [40] who found that it was not possible to obtain a closed form solution for the parameters of the all-pole model except for the case in which the pitch period was much greater than the model order. El-Jaroudi and Makhoul [41] also addressed this problem and derived a set of nonlinear equations whose solution also required use of iterative techniques. It will now be shown that these problems can be avoided if the all-pole model is fitted to the cubic spline envelope rather than to the sine-wave amplitudes.

5.2. Computation of the parameters of the all-pole model

The next step is to develop an analytical method for computing the parameters of the all-pole model such that its magnitude is a best fit to the spline approximation to the vocal tract envelope. Letting $H_a(\omega)$ represent the transfer function of the all-pole model then

$$H_a(\omega) = \frac{\sigma}{A(\omega; \mathbf{a})}$$

$$A(\omega; \mathbf{a}) = 1 - \sum_{k=1}^{p} a_k \exp(-jk\omega) \tag{5.1}$$

where σ and $\mathbf{a} = (a_1, a_2, ..., a_p)$ are the parameters to estimated. Since the ear responds logarithmically to the sine-wave amplitudes, a criterion that is well-suited to the speech application is to minimize the average dB error. Letting

$$E(\omega) = log|H_s(\omega)|^2 - log|H_a(\omega)|^2 \tag{5.2}$$

denote the error between the measured and the modeled sine-wave amplitude envelopes, then a reasonable approach might be to pick σ and $\mathbf{a}$ to minimize the average error

$$\epsilon(\sigma, \mathbf{a}) = \frac{1}{\pi} \int_0^{\pi} E^2(\omega) d\omega \tag{5.3}$$

Unfortunately it is not possible to obtain an analytically tractable closed-form expression for the optimizing values of the all-pole parameters. However if the

symmetric squared-error criterion is modified to allow for an asymmetry of the form,

$$\epsilon(\sigma, \mathbf{a}) = \frac{1}{\pi}\int_0^\pi [\exp E(\omega) - E(\omega) - 1]d\omega \tag{5.4}$$

then a closed-form solution is possible. Using eq. (5.1) in eq. (5.2) the logarithmic error can be written as

$$E(\omega) = \log\left[\frac{|H_s(\omega)A(\omega)|^2}{\sigma^2}\right] \tag{5.5}$$

and then the error criterion in eq. (5.4) becomes

$$\epsilon(\sigma, \mathbf{a}) = \frac{1}{\pi}\int_0^\pi \left[\frac{|H_s(\omega)A(\omega)|^2}{\sigma^2} - \log|H_s(\omega)A(\omega)|^2 + \log\sigma^2 - 1\right]d\omega \tag{5.6}$$

By differentiating eq. (5.6) with respect to σ it is easy to show that its optimum value, $\hat{\sigma}$, is given by

$$\hat{\sigma}^2 = \frac{1}{\pi}\int_0^\pi |H_s(\omega)A(\omega)|^2 d\omega \tag{5.7}$$

and if this optimizing value is then substituted back into eq. (5.6) the resulting error is given by

$$\epsilon(\mathbf{a}) = \frac{1}{\pi}\int_0^\pi |H_s(\omega)A(\omega)|^2 d\omega \tag{5.8}$$

If this expression is differentiated with respect to each of the coefficients, a_k, it is also easy to show that their optimum values, $\hat{a}_k$, satisfy the equations

$$\sum_{k=1}^{p} R_{j-k}\hat{a}_k = R_j \quad j = 1, ..., p$$

$$R_k = \frac{1}{\pi}\int_0^\pi |H_s(\omega)|^2 \cos(k\omega)d\omega \tag{5.9}$$

The latter set of equations can be recognized as the normal equations that arise in the time-domain solution for the linear predictor coefficients where in that case, R_k represents the kth time-domain correlation coefficient. By performing the analysis using frequency-domain data, the R_k in eq. (5.9) can also be interpreted as correlation coefficients, but they are now computed from the cubic spline fit to the magnitude of the vocal tract transfer function.

Of course it will have been recognized that the above solution is simply a generalization of the analysis developed by Itakura and Saito [42] for finding the maximum likelihood estimate of the power spectral density for Gaussian processes and that the error metric in eq. (5.4) is the so-called Itakura-Saito spectral matching criterion [43]. The above frequency-domain analysis has several advantages: it applies both to voiced and unvoiced speech, it removes the dependency on pitch and it allows for the use of very high order models, which, in the limit, can track the spline envelope exactly. Moreover it gives a precise expression for the gain, σ, which applies regardless of the voicing state. Such a "clean" definition of the gain has always proved elusive in the time-domain analysis since the application of the autoregressive model to periodic voiced speech is not theoretically correct. Moreover, the estimated parameters, $\hat{\mathbf{a}}$, are functionally the same as the predictor coefficients that arise in the time-domain analysis, hence all of the techniques used in linear prediction analysis can be applied to the parameters of the all-pole model. For example, the matrix equation in eq. (5.9) is Toeplitz, hence its solution can be readily obtained using the Levinson-Durbin algorithm. Finally, the alternate representations for the frequency-domain "predictor coefficients", such as the reflection coefficients or the line spectral frequencies, can also be applied to these all-pole parameters so that the scalar and vector quantization techniques conventionally employed in LPC- or CELP-based systems can be applied to efficiently encode the parameters of the all-pole model. This topic will be addressed in more detail later in this section. The result of fitting 10'th and 22'nd order all-pole models to the cubic spline envelope are shown in fig 8. A sine-wave synthesis system was developed for all-pole models of arbitrary order by computing a set of sine-wave amplitudes and a set of sine-wave system phases by taking the harmonic samples of the magnitude and phase of the reconstructed all-pole transfer function. In conjunction with the pitch-dependent, linear-phase model and the voicing-adaptive random phases, the quality of the synthetic speech was quite good, in many cases almost equivalent to the original minimum phase system provided the model order was $p \geq 22$. By comparison, if the model order was set to be $p = 10$, then the synthetic speech was mechanical and buzzy, having the same general quality of a basic 10'th order LPC system. This shows that the phase randomization controlled by the estimate of the voicing probability is, in itself, insufficient to capture all of the voicing information implicit in the sine-wave data. In fact it appears that there is considerable voicing information contained in the sine-wave amplitudes which is most effectively transferred to the synthetic speech through the phase of the all-pole model. More precisely it is the time-rate-of-change of the system phase that imparts a more natural sound to the synthetic speech and is a major consideration in keeping the frame-rate for the all-pole model high, as close to about 15 ms as possible. Although it is possible to code a high-order all-pole model adequately at 4800 b/s, encoding such a system at 2400 b/s is difficult since the number of bits per parameter would be too low to ensure good quality speech. Therefore, it is desirable to reduce the model order as much as possible without introducing the mechanical, buzzy quality. This can be done using the notion of spectral warping, a topic that will be discussed in the next section.

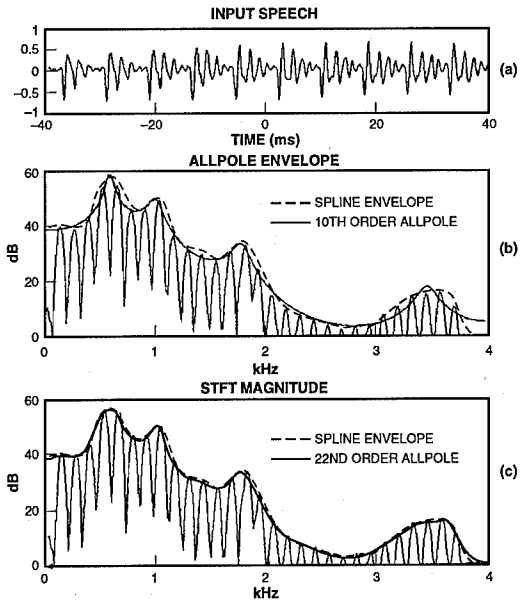


Figure 8. Comparison of 10'th and 22'nd order allpole fits to the spline envelope.

5.3. Spectral warping

It is well-known that the ear is less sensitive to details in the sine-wave amplitudes at higher frequencies than at lower frequencies, a property which has been exploited extensively in the past by frequency-domain coders such as the phase vocoder and the channel vocoder [44]. Generally the bandwidth of the filters in these systems is increased logarithmically to maintain a constant Q, similar to the cochlear filters that approximate the front-end of the hearing mechanism. These so-called critical bandwidths can be approximated analytically using the bark scale or the mel scale. If ω' represents the perceptual scale, then the relationship between the perceptual scale and the frequency scale can be written as $\omega' = W(\omega)$ for some warping function W. One way to exploit the perceptual warping property is to map the measured vocal tract envelope onto the perceptual scale and then fit the all-pole model to this warped envelope. The sine-wave amplitude and system phase at a given frequency ω can then be found by sampling the magnitude and phase of the warped all-pole transfer function on the perceptual scale at $W(\omega)$.

Although the standard perceptually-based warping functions, such as the mel scale or the bark scale, could be used, a more general procedure was developed to allow for more flexibility in designing the coder at a multiplicity of bit rates. In particular, it is desired to maintain a linear mapping at least within the region

of the first formant ($\geq$ 800 Hz) with a logarithmic mapping in the high-frequency region, a relationship which can be described parametrically as

$$W(\omega) = \alpha \log(1 + \beta\omega) \tag{5.10}$$

A warping function that has been found to work well for the 2400 b/s and 4800 b/s systems uses $\alpha = 170$ and $\beta = .554$. A typical example of a cubic spline envelope fitted to the measured sine-wave amplitudes and plotted on the linear frequency scale is shown in fig. 9(b). Fig. 9(c) shows the result of plotting the cubic spline envelope on the warped scale. An example of fitting a $14'th$ order all-pole model to the warped spline envelope is shown in fig. 9(c). The envelope of the all-pole model that was fitted on the warped scale but plotted on the original frequency scale is shown in fig. 9(d). Usually there is some loss of detail at high frequencies. When

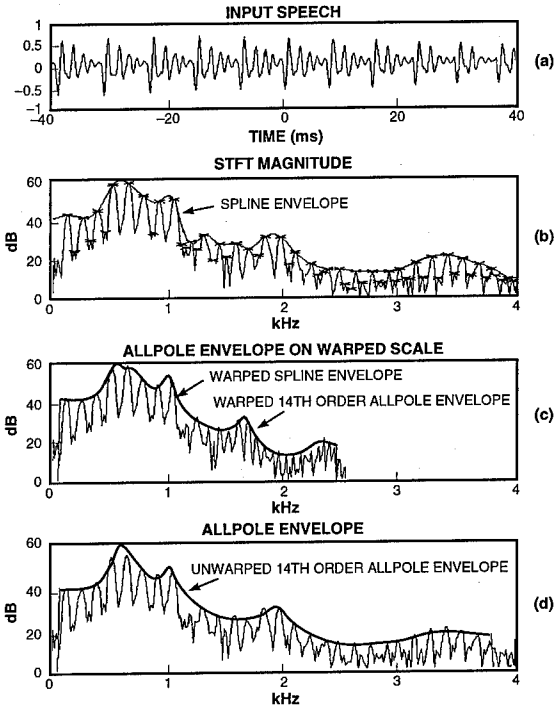


Figure 9. Example of a 14'th order allpole fit to the warped spline envelope.

the $14'th$ order warped all-pole model was used in place of the cubic spline envelope in the minimum-phase harmonic sine-wave synthesis system, little degradation in the quality of the synthetic speech was perceived. If no warping was used, the

synthetic speech took on a mechanical constrained quality which was particularly pronounced for certain speakers. On the other hand, there were some speakers for which no difference could be perceived. The next step is to develop the quantization rules for encoding the parameters of the warped all-pole model, a topic which will now be addressed.

5.4. Quantization of the parameters of the all-pole model

It has been established that synthetic speech of good quality is possible by encoding the sine-wave amplitudes in terms of the parameters of the all-pole model. Therefore good digital speech communications is possible provided the gain, $\hat{\sigma}$ and the "predictor coefficients" $\hat{a}_1, \hat{a}_2, ..., \hat{a}_p$ can be efficiently quantized. One way to encode the gain is to quantize $\log \hat{\sigma}$ uniformly over about a 60-90 dB dynamic range. Allowing for a 1 dB quantization noise, this requires about 6 bits per frame update. Since it is desirable to operate at a 15 ms frame-rate, this quantization scheme, while acceptable at the 4800 b/s rate, expends too many bits at the 2400 b/s rate. In the latter case it has been found useful to use a simple predictor and quantize the prediction residual. If $g(k) = \log \hat{\sigma}$ is the measured gain for the k'th frame and if $\tilde{g}(k-1)$ is the decoded gain for the $(k-1)$'st frame, then the decoded gain for frame k, $\tilde{g}(k)$ is

$$\tilde{g}(k) = \alpha \tilde{g}(k-1) + Q[g(k) - \alpha \tilde{g}(k-1)] \tag{5.11}$$

where α is the gain prediction coefficient ($\approx .7$). and $Q[\cdot]$ represents an appropriately designed nonlinear quantization table. This method has been found to give good results using about 4 bits to quantize the residual gain.

The next step is to quantize the "predictor coefficients", but as has been found in time-domain LPC it is better to work with a transformation of these coefficients rather than try to encode them directly. Drawing upon the significant body of research in this area the "predictor coefficients" are transformed to the line spectral frequencies (LSFs) and these can be either scalar-quantized or vector-quantized using standard methods. However, due to the complexity of implementing a vector-quantized set of LSFs for the higher-order all-pole models, the focus has been mainly to exploit the properties of the differential LSFs as described by Soong and Juang [45]. Their design methods have been applied to LSF data gathered for a number of model orders for all-pole models that were fitted on the warped frequency scale.[2].

Since it typically takes about 3 bits per LSF on average, a $14'th$ order all-pole model would require about 42 bits to achieve a reasonable level of performance. While it is not a problem to encode this many bits at 4800 b/s, at 2400 b/s it requires that a frame rate of about 30 ms be used. In all of the experiments reported

[2] The authors are indebted to J.L.Trent and T.G.Champion for their work in developing the design software and in generating a multitude of quantization tables

in this chapter, good quality synthetic speech required that the frame rate be about 15 ms. Therefore there arises a basic incompatibility in encoding the all-pole model at the lower data rate. While it is possible to devise means for encoding the pitch, voicing and gain for operation at this data rate using the 15 ms frame rate, it is obvious that some additional property needs to be exploited to encode the shape of the all-pole spectrum with the required temporal fidelity. One way to approximate the desired temporal resolution is to use the McLarnen frame-fill algorithm [46].

5.5. *Frame-fill interpolation*

A simple and straightforward approach to reducing the transmission bandwidth is to transmit the spectral data for every second frame, using some control information to instruct the synthesizer how to reconstruct (or "fill-in") the missing information. A set of rules that has been found to be quite useful is simply to compare the spectral data for the frame to be omitted to the quantized spectral data on the preceding and succeeding frames. The data that "best" represents the mid-frame spectral shape is determined, and one bit can be used to signal the decision to the receiver. An additional bit can be used to allow for candidate fits that are interpolated values of the end-point data. Therefore, in the present context two bits are used, allowing for four frame-fill options [46].

If $\tilde{S}_{k-1}(\omega)$ and $\tilde{S}_{k+1}(\omega)$ represent the quantized gain-normalized spectral shapes on frame $k-1$ and frame $k+1$, then the four frame-fill options are:

$$\begin{aligned} log S_k^1(\omega) &= log \tilde{S}_{k-1}(\omega) \\ log S_k^2(\omega) &= log \tilde{S}_{k+1}(\omega) \\ log S_k^3(\omega) &= .5\, log \tilde{S}_{k-1}(\omega) + .5\, log \tilde{S}_{k+1}(\omega) \\ log S_k^4(\omega) &= .333\, log \tilde{S}_{k-1}(\omega) + .667\, log \tilde{S}_{k+1}(\omega) \end{aligned} \tag{5.12}$$

If $S_k(\omega)$ represents the unquantized gain-normalized spectral shape at frame k, then a reasonable measure of the error when each of the frame-fill options is used is given by the squared-log-error criterion:

$$d(S_k^i, S_k) = \frac{1}{\pi} \int_0^\pi [log S_k^i(\omega) - log S_k(\omega)]^2 d\omega \qquad i = 1, 2, 3, 4 \tag{5.13}$$

and the frame-fill spectral shape, $\tilde{S}_k(\omega)$ is determined for the condition resulting in the smallest value of $d()$. Ideally the frame-fill decision should be made comparing the actual vocal tract shapes, but as the computation of the envelope of the all-pole model adds considerable complexity to a practical implementation, an approximate spectral decision criterion is used. Since the LSFs evolve relatively smoothly in time, it is not unreasonable to base the frame-fill options on interpolated values of the quantized LSFs. Letting $\omega = (\omega_1, \omega_2, ..., \omega_p)$ denote the vector of LSFs, and letting $\tilde{\omega}_{k-1}$ and $\tilde{\omega}_{k+1}$ denote the quantized LSFs on frames $k-1$ and $k+1$ respectively,

then the four frame-fill options can be written as

$$\begin{aligned}\omega_k^1 &= \tilde{\omega}_{k-1} \\ \omega_k^2 &= \tilde{\omega}_{k+1} \\ \omega_k^3 &= .5\,\tilde{\omega}_{k-1} + .5\,\tilde{\omega}_{k+1} \\ \omega_k^3 &= .333\,\tilde{\omega}_{k-1} + .667\,\tilde{\omega}_{k+1}\end{aligned} \tag{5.14}$$

Drawing on the work of Paliwal and Atal [47] in the context of selecting the best codeword for an LSF-based vector quantizer, a reasonable criterion for picking the best frame-fill option is the minimze the weighted squared LSF-error. If $\bar{\omega}_k = (\bar{\omega}_1, \bar{\omega}_2, \cdots, \bar{\omega}_p)$ represents the unquantized LSF vector at frame k, then the error in using each of the frame-fill vectors can be measured using the criterion

$$d(\omega_k^i, \bar{\omega}_k) = \sum_{m=1}^{p} W(\omega_m)(\omega_m^i - \bar{\omega}_m)^2 \tag{5.15}$$

where the weighting term is computed by sampling a compressed value of the cubic spline envelope at each of the measured frame-k LSF frequencies. Since the LSFs represent the all-pole model on the warped scale, the weighting must be computed for the spline envelope that is also on the warped scale. Therefore, the weights are

$$W(\omega_m) = |H_s[W(\omega_m)]|^\gamma \tag{5.16}$$

where the compression factor, $\gamma \approx .5$.

In the above application of the frame-fill algorithm the four options were applied over the entire speech bandwidth. At low data rates where only two bits might be available for the frame-fill option, this may be the only feasible alternative for estimating the mid-frame spectral shape. However if two additional bits could be made available through more efficient spectral coding, for example, then the speech bandwidth could be split in half and two sets of frame-fill decisions could be used. This idea can be extended to allow for a multiplicity of spectral bands and this would allow for better tracking of the individual formant bandwidths. An extreme case was to apply the frame-fill option to each of the p LSFs. In this test the synthetic speech had the same lively, natural quality that was obtained using the unquantized spectral shape at the 15 ms frame rate.

5.6. Experimental results

Using the techniques described in the previous sections low-rate coders were developed at 4800 b/s and 2400 b/s data rates using measurements of the pitch, voicing, gain, and LSFs that were made every 15 ms. At 4800 b/s the pitch (7 bits), voicing (3 bits) and gain (6 bits) were encoded at the 15 ms rate, but the LSFs were encoded only once per 30 ms outer frame using multiband frame-fill to provide

the amplitude information at the inner frame rate. For a $14th$ order system, the frame-fill technique was applied to each of the LSFs using 28 bits. The remaining bits were used to encode the frequency-differential LSFs using scalar quantization.

At 2400 b/s the pitch was coded on the 30 ms outer frame using 6 bits, and one frame-fill bit was used to encode the pitch on the inner frame. The voicing probability was encoded on the 15 ms inner frame with 2 bits per inner frame. The gain was coded differentially on the 15 ms inner frame using 4 bits per frame. Only 2 bits were assigned to frame-fill the LSFs and the remaining bits were used to encode the frequency-differential LSFs using scalar quantization.

Since the overlap-add technique is used to perform the sine-wave speech synthesis, the 15 ms inner frame size would correspond to a 30 ms interpolation window. This is too wide to give good perceptual results. Therefore the synthesizer framing was further divided into two 7.5 ms sub-frames per 15 ms inner frames. The sine-wave parameters for the sub-frames were obtained by interpolating the pitch, voicing, gain and amplitude envelopes obtained for the 15 ms inner frame. It is important to note that the voicing-adaptive phase model was applied at the 7.5 ms sub-frame rate as this allowed for more randomization of the unvoiced phases which also results in synthetic speech that was more pleasant to listen to. The above coder has come to be referred to as the *sinusoidal transform coder* (STC).

The 2400 b/s version of STC is one of the vocoders to be evaluated in September 1995 as a candidate for a new U.S. Government standard to replace the LPC10e algorithm. The system was evaluated in a pre-selection test in June of 1994 in comparison with Federal Standard 1016 CELP algorithm at 4800 b/s. Its MOS, DAM and DRT scores are given in table 1 [48]

Table 1.

condition	STC 2400 b/s	CELP 4800 b/s
quiet	3.525 MOS	3.592 MOS
office noise	3.021 MOS	2.938 MOS
quiet	65.3 DAM	63.1 DAM
quiet	90.9 DRT	92.8 DRT

There is, at this time, no similar comparisons of the performance of the 4800 b/s STC to either the CELP coder or to STC at 2400 b/s. However, in 1992, the 4800 b/s cepstral-based version of STC was evaluated in the TIA Half-Rate Digital Cellular preselection test and its performance, along with two other coders, was deemed to belong to the same equivalence class as the 8000 b/s VSELP algorithm [49]. Currently the major focus of the 4800 b/s system is in its application to

the development of an appliqué for the Secure Telephone Unit (STU-III) for secure, multi-speaker conferencing. This topic will now be discussed.

5.7. Multi-speaker conferencing

One of the most important features in facilitating speech conferencing is to allow for speaker interruption as then the control of the conference can move naturally among all conferees. In analog or wide-band digital conferencing a speaker interrupt, which corresponds to two speakers talking simultaneously, is handled by signal summation at a conferencing bridge. Such a scheme is not possible for digital speech coders as these would require synthesis and reanalysis of the aggregate speech signal, a process called tandeming which almost always results in a severe loss in quality even if only a single speaker is talking. Further degradations occur during a speaker interrupt since most low-rate coders are designed to model only a single voice.

A technique that has been developed collaboratively amongst Rome Laboratories, Lincoln Laboratories and ARCON Inc., defers signal summation to the synthesis terminal by adaptively allocating the available bandwidth based on the number of active talkers [50]. Since during most of the conference there will only be a single speaker talking at any one time, the quality of the speech will be maintained at the highest level since the single speaker is always encoded at the highest rate. When there are two speakers present, each speaker is allocated one-half of the bandwidth, and although the quality of the individual speakers will be somewhat reduced, intelligibility of the two speakers will be preserved. Overall there will be a more natural contention for the conference control.

A critical component in this conferencing system is in the bridge which controls signal routing and bit-rate reduction on those parameter sets when a two-speaker overlap has occurred. When there is only one active speaker, all conferees (except the active speaker) receive the same set of parameters at the highest rate. When there are two active speakers, each speaker receives the other speaker's parameters at the highest rate, while the passive listeners receive the parameters sets of the two active speakers, each transformed to the lower rate.

It is the dynamic multirate capability of STC that lends itself naturally to the implementation of this transformation process. Since the coder depends only on pitch, voicing, gain and LSFs it is not necessary to synthesize the speech and repeat the analysis for operation at the lower rate. Rather the parameters are decoded at the higher rate and then encoded using quantization tables that were designed for operation at the lower rate. Moreover, the use of STC makes possible a simple technique for implementing the double-speaker signal summation since the summation can be done in the sine-wave domain prior to synthesis, greatly reducing the computational complexity of the synthesizer. Using the overlap-add sine-wave synthesis technique the pitch, voicing and spectral parameters are used to compute the amplitudes, frequencies and phases of the underlying sine waves, and these are used to fill the complex DFT buffer at the sine-wave frequencies. The speech waveform is obtained from the inverse DFT. With two speakers, the two sets of complex

parameters are added in the complex transform domain before taking the inverse transform. In this way the synthesis of two speakers involves only slightly more computation than for one speaker.

At Rome Laboratories a conferencing system based on the above signal processing ideas has been implemented in real-time. The bridge can handle up to four conferees with two speakers active at any one time using 4800 b/s at the highest rate for a single speaker and 2400 b/s for the two-speaker overlap. The system has been tested extensively using both quiet and noisy conditions that include the Airborne Commmand Post and the F15 fighter cockpit environments. It has been found to provide digital speech of very high quality and allows for a natural flow of the conference control while being robust under adverse noise conditions.

6. Improved multi-band excitation vocoder

One of the more successful applications of the sine-wave modeling technique to low-rate speech coding is the *improved multi-band excitation* (IMBE) speech coder. Versions of this coder have been chosen as standards for the INMARSAT-M system in 1990 [51], the APCO/NASTD/Fed Project 25 in 1992 [52] and the INMARSAT-Mini-M system in 1994 [53]. The latter system demonstrated performance at a gross rate of 4800 b/s that was at least as good as that achieved by the full-rate digital IS-54 standard VSELP algorithm operating at 8000 b/s in a multitude of categories including acoustic noise and channel errors. It is currently under consideration for a number of emerging digital communications systems including PCS and cellular radio. Since detailed description of the techniques used in IMBE are readily available in the Inmarsat and APCO specifications as well as in the recent book by Kondoz [14] the discussion in this section will be more theoretical showing the similarities and differences of IMBE versus the generic sine-wave coding methods described earlier in this chapter.

6.1. Harmonic sine-wave model

The starting point for the original *multi-band excitation* (MBE) speech coder developed by Griffin and Lim [10] was to represent speech as a sum of harmonic sine waves. As in eq. (3.2) in section 3, the synthetic waveform for a harmonic set of sine waves is written as

$$\hat{s}(n) = \sum_{\ell=1}^{L(\omega_0)} A_\ell \exp(jn\ell\omega_0 + \phi_\ell) \tag{6.1}$$

Whereas in section 3.2 the sine-wave amplitudes were assumed to be harmonic samples of an underlying vocal tract envelope, in MBE they are allowed to be unconstrained free variables and are chosen to render $\hat{s}(n)$ a minimum-mean-squared-error fit to the measured speech signal $s(n)$. In MBE the error is measured using

windowed versions of the speech signals where the windowing functions are more general than the rectangular window that was used in section 3. Letting the window function be $w(n)$, which typically might be a Hamming window or a Kaiser window, the weighted speech signals are $s_w(n) = w(n)s(n)$ and $\hat{s}_w(n) = w(n)\hat{s}(n)$ and the mean-squared-error is given by

$$\epsilon(\omega 0, \mathbf{A}, \phi) = \sum_{-N}^{N} \mid s_w(n) - \hat{s}_w(n) \mid^2 \tag{6.2}$$

where $\mathbf{A} = (A_1, A_2, \cdots, A_{L(\omega_0)})$ and $\phi = (\phi_1, \phi_2, \cdots, \phi_{L(\omega_0)})$ are the vectors of unknown amplitudes and phases at the sine-wave harmonics. If

$$S_w(\omega) = \int_{-\pi}^{\pi} s_w(n) \exp(-jn\omega) d\omega \tag{6.3}$$

denotes the discrete-time Fourier transform of $s_w(n)$ and $\hat{S}_w(\omega)$ denotes the discrete-time Fourier transform of $\hat{s}_w(n)$, then using Parseval's theorem, eq. (6.2) becomes

$$\begin{aligned}\epsilon(\omega_0, \mathbf{A}, \phi) &= \int_{-\pi}^{\pi} \mid S_w(\omega) - \hat{S}_w(\omega) \mid^2 d\omega \\ &= \int_{-\pi}^{\pi} \left\{ \mid S_w(\omega) \mid^2 - 2Re\left[S_w(\omega)\hat{S}_w^*(\omega)\right] + \mid \hat{S}_w(\omega) \mid^2 \right\} d\omega \end{aligned} \tag{6.4}$$

The first term, which is independent of the pitch, amplitude, and phase parameters, is the energy in the windowed speech signal, E_w. Letting $\alpha_\ell = A_\ell \exp(j\phi_\ell)$ represent the complex amplitude of the ℓth harmonic and using the sine-wave decomposition in eq. (6.1), $\hat{S}_w(\omega)$ can be written as

$$\hat{S}_w(\omega) = \sum_{\ell=1}^{L(\omega_0)} \alpha_\ell W(\omega - \ell\omega_0) \tag{6.5}$$

where $W(\omega)$ is the Discrete-time Fourier Transform of the windowing function $w(n)$. Substituting this relation into eq. (6.4), the mean-squared-error can be written as

$$\begin{aligned}\epsilon(\omega_0, \mathbf{A}, \phi) = E_w &- 2Re \sum_{\ell=1}^{L(\omega_0)} \alpha_\ell^* \int_{-\pi}^{\pi} S_w(\omega) W(\omega) d\omega \\ &+ \sum_{\ell=1}^{L(\omega_0)} \sum_{m=1}^{L(\omega_0)} \alpha_\ell \alpha_m^* \int_{-\pi}^{\pi} W(\omega - \ell\omega_0) W(\omega - m\omega_0) d\omega \end{aligned} \tag{6.6}$$

Since for each value of ω_0 this equation is quadratic in α_ℓ, it is straightforward to solve for the $\alpha(\omega_0)$ that results in the minimum the mean-squared error, $\epsilon(\omega_0, bf\alpha(\omega_0))$. This process can be repeated for each value of ω_0 such that the optimum minimum-mean-squared-error estimate of the pitch can be determined. Although the quadratic optimization problem is straigh-forward to solve, it requires solution of a simultaneous set of linear equations that have to be solved for each candidate pitch value. This renders the resulting pitch estimator complex to implement. However, following [10], if it is assumed that $W(\omega)$ is essentially zero in the region $|\omega| > \omega_0/2$, which corresponds to the condition posed in section 3 to insure that the sine-waves are well-resolved, and if

$$\Omega_\ell = \{\omega : \ell\omega_0 - \omega_0/2 \leq \omega \leq \ell\omega_0 + \omega/2\} \tag{6.7}$$

then the mean-squared-error can be approximated as

$$\begin{aligned}\epsilon(\omega_0, \mathbf{A}, \phi) = E_w - 2Re \sum_{\ell=1}^{L(\omega_0)} \alpha_\ell^* \int_{\Omega_\ell} S_w(\omega) W(\omega) d\omega \\ + \sum_{\ell=1}^{L(\omega_0)} |\alpha_\ell|^2 \int_{\Omega_\ell} W^2(\omega - \ell\omega_0) d\omega\end{aligned} \tag{6.8}$$

from which it follows that the value of the complex amplitude that minimizes the mean-square-error is

$$\hat{\alpha}_\ell(\omega_0) = \frac{\int_{\Omega_\ell} S_w(\omega) W(\omega - \ell\omega_0) d\omega}{\int_{\Omega_\ell} W^2(\omega - \ell\omega_0) d\omega} \tag{6.9}$$

The best mean-squared-error fit to the windowed speech data is then given by

$$\hat{S}_w(\omega; \omega_0) = \sum_{\ell=1}^{L(\omega_0)} \hat{\alpha}_\ell(\omega_0) W(\omega - \ell\omega_0) \tag{6.10}$$

where the dependence on the specified pitch, ω_0, is now made explicit. This expression is then used in eq. (6.2) to evaluate the mean-squared-error for the given value of ω_0. This procedure is then repeated for each value of ω_0 in the pitch range of interest and the optimum estimate of the pitch is the value of ω_0 that minimizes the mean-squared-error.

While the procedure is similar to that used in section 3, there are important differences. For one thing, the use of the unconstrained amplitude estimates will render the MBE pitch estimator ambiguous at the multiples of the pitch period and heuristic methods must be used to remove the ambiguity. For another, the procedure depends on the Discrete-time Fourier Transform of the windowed speech signals

and hence it is not possible to introduce perceptually-based amplitude compression that is often needed to compensate for the large dynamic range of neighboring sine waves. Finally, because the estimator depends on the phase of the Discrete-time Fourier Transform over each harmonic lobe, the resulting amplitude estimates can often be meaningless. In fact in the MBE systems, the above amplitude estimates are *not* used to model the sine-wave amplitudes but only to estimate the pitch and voicing. The errors in the amplitudes are probably due to the fact that the phase of the Discrete-time Fourier Transform is not always constant across every harmonic lobe as is assumed in the above formulation of the estimation problem.

While the implementation of the MBE pitch estimation algorithm is intensive, even after the approximations introduced in the above analysis, modifications have been developed to reduce the complexity significantly [14]. In practice a two-step procedure is used in which baseband speech out to about 1000 Hz is used in a correlation pitch estimator using a 37.5 ms Kaiser window to develop a coarse estimate of the pitch. Forward-backward pitch tracking is used to maintain a meaningful pitch track during regions in which the pitch and/or vocal tract are rapidly changing. Then a 27.5 ms Kaiser window is applied to full-band speech data and the above algorithm is used to refine the pitch estimate according to equations 6.9-6.10.

6.2. Multi-band voicing

The next step in the MBE algorithm is to distinguish between regions of voiced and unvoiced sine waves. As was done in section 3 this is based on how well the harmonic set of sine waves fits the measured set of sine waves. In section 3 a signal-to-noise ratio (SNR) was defined in terms of the normalized mean-squared-error, eq. (3.27) and this was mapped into a cutoff frequency below which the sine waves were declared voiced and above which they were declared unvoiced. This idea, which originated with the work of Makhoul et al. [34], was generalized by Griffin and Lim [10] to allow for an arbitrary sequence of voiced and unvoiced bands with the measure of voicing in each of the bands determined by the normalized mean-squared-error computed for the windowed speech signals. Letting

$$B_m = \{\omega : \omega_{m-1} \leq \omega \leq \omega_m\} \quad m = 1, 2, \cdots, M \tag{6.11}$$

denote the *mth* band of a multiband of M regions of the speech bandwidth, then using eq. (6.4), the normalizing mean-squared-error for each band can be written as

$$\epsilon_m(\hat{\omega}_0) = \frac{\int_{B_m} \mid S_w(\omega) - \hat{S}_w(\omega; \hat{\omega}_0) \mid^2}{\int_{B_m} \mid S_w(\omega) \mid^2} \tag{6.12}$$

Each of the M values of the normalized mean-squared-error is compared with a threshold function to determine the binary voicing state of the sine waves in each

band. If $\epsilon_m(\hat{\omega}_0)$ is below the threshold, the mean-squared error is small, hence the harmonic sine waves fit well to the input speech and the band is declared voiced. The setting of the threshold is very important and MBE uses several heuristic rules to get the best performance. The most recent set of rules given in [14] shows the threshold decreasing with frequency, decreasing if on the previous frame the band was unvoiced, decreasing if the high-frequency energy exceeds the low-frequency energy, and decreasing if the speech energy approaches the energy of the background noise. In other words, every effort is made to favor an unvoiced declaration, an observation that has been confirmed by experiments that show that without the high-low energy measure and the comparison against the background noise level, the synthetic speech produced by MBE is very buzzy. In another set of experiments, the multiband voicing decisions were combined into a two-band voicing-adaptive cutoff frequency as was used in section 4 and no loss is quality was perceived. This was confirmed by an independent set of experiments reported in the literature [14, 54]. Apparently the advantage of multiband voicing lies not with its ability to mix voiced and unvoiced states throughout the speech bandwidth, but rather to make reliable voicing decisions when the speech signal has been corrupted by additive acoustical noise and deleterious spectral shaping such as when an IRS filter is used. The reason for this lies in the fact that the normalized mean-squared-error essentially removes the effect of the spectral tilt which means that the sine-wave amplitudes contribute more or less equally from band to band. When one wide-band voicing decision is made, only the largest sine-wave amplitudes will contribute to the mean-squared-error, and if these have been corrupted due to IRS filtering or noise, then the remaining sine-waves, although harmonic, may not contribute enough to the error measure to offset those that have been altered by the front end processing.

6.3. Sine-wave amplitude model

Since the amplitudes that were computed during the pitch estimation process prove to be poor representations of the underlying sine-wave amplitudes, IMBE uses a different method to estimate them. Since the pitch has been determined, the model for the Discrete-time Fourier Transform of the harmonic sine-wave speech model within the region Ω_ℓ of the ℓth harmonic can be written explicitly as

$$\hat{S}_w(\hat{\omega}_0) = A_\ell \exp(j\phi_\ell) W(\omega - \ell\hat{\omega}_0) \quad for\ \omega\ in\ \Omega_\ell \tag{6.13}$$

In MBE the amplitude is chosen such that over each harmonic region, Ω_ℓ, the energy in the model matches in the measured energy. This leads to the amplitude estimator

$$\hat{A}_\ell = \left[\frac{\int_{\Omega_\ell} |\, S_w(\omega)\,|^2\, d\omega}{\int_{\Omega_0} W^2(\omega) d\omega} \right]^{\frac{1}{2}} \tag{6.14}$$

An additional processing step is required before the sine-wave amplitude data is presented to the synthesizer. In the later versions of MBE [14] this is referred to as "Spectral Amplitude Enhancement" which is another term for the post-filtering operation that is used in almost all contemporary low-rate coders. In fact the method used in the MBE systems is based on the frequency-domain design principles introduced by McAulay and Quatieri [37] and described in detail in in section 4. If the post-filter weights in eq. (4.23) are evaluated using a compression factor $\gamma = .5$, they become

$$W_\ell = \sqrt{\hat{A}_\ell}\left[\frac{\pi[R_0^2 - 2R_1R_0\cos(\ell\omega_0) + R_1^2]}{\hat{\omega}_0 R_0(R_0^2 - R_1^2)}\right]^{\frac{1}{4}} \tag{6.15}$$

where use has been made of the fact that $L = \pi/\hat{\omega}_0$ is the number of harmonics in the speech bandwidth. These can be identified as the same weights given in the Inmarsat [51] and APCO [52] specifications. The idea of clipping the post-filter weights, eq. (4.24), which was first introduced in the Inmarsat system and later refined in the APCO system, has proven to be an useful modification to the frequency-domain post-filtering technique as it allows for a more relaxed compression factor which in turn sharpens the formant bandwidths leading to synthetic speech which is much crisper.

6.4. Sine-wave phase model and voiced-speech synthesis

In MBE two distinct methods are used to synthesize voiced and unvoiced speech. Basically voiced speech is generated using the overlap-add method described in section 3 with the exception that if the pitch on successive frames does not change by more than 10% then the first eight sine-waves will be matched and the phases will be interpolated using a function that maintains phase continuity at the frame boundaries at the expense of a slight discontinuity in the frequencies. This is a simplification of the cubic phase interpolation technique that was described in [3, 8]. Apparently the latter condition is needed to allow the MBE synthesizer to run at a fixed 20 ms frame rate since otherwise the sine-wave parameters can become non-stationary over the duration of the overlap-add window, an effect that seems to be particularly important for some high-pitched speakers. This problem is avoided in STC at the expense of adding some complexity to the synthesizer by interpolating the the sine-wave parameters to $\approx$ 10 ms frame size and then repeating the overlap-add procedure at the faster frame rate.

For the all of the voiced sinewaves it is necessary to specify the phases at the frame boundaries. In the MBE the phases were coded differentially in time and even at 8 kb/s not all the phases could be coded for some low-pitched speakers. The uncoded phases were made random and this gave the synthetic speech a reverberant quality. Although improvements in the efficiency of the phase coder were made, arbitrary phase randomization was needed for some of the phases for some low-pitched speakers and reverberation continued to be a problem [56]. Then after the

sine-wave phase model was described in [3] and again in [37], the coherent excitation phase model was incorporated into IMBE [51, 52] and the reverberation problem was eliminated. Therefore the synthetic phase model used in the latest versions of IMBE use the excitation phase model described in section 4 in which the phase of the fundamental is the integral of the instantaneous pitch frequency and then the phase of the ℓth harmonic is ℓ times the phase of the fundamental.

It should be noted that neither the Inmarsat nor the APCO versions of IMBE make use of the vocal tract phase, using instead a simple zero-phase model. In STC the minimum-phase phase is essential as it adds more naturalness and crispness to the synthetic speech. Whether the addition of the minimum-phase phase to the IMBE synthesis system would improve the quality of the synthetic speech in the same way is, as yet, an open question.

6.5. Unvoiced synthesis

For those speech bands for which the sine waves have been declared unvoiced, MBE synthesis is done using filtered white noise. Care needs to be taken to insure that the effects of the analysis and synthesis windows have been removed so that the correct synthesis noise level is achieved. The details of the normalization procedures are given in [14]. This approach to unvoiced synthesis is in contrast to STC which uses random phases in the unvoiced regions. The advantage of using random phases is that the synthesizer is simpler to implement as exactly the same operations are performed for voiced and unvoiced speech. This happens to be particularly advantageous in applications that exploit the multirate capabilities of STC [50, 58].

6.6. Sine-wave parameter coding

In order to operate MBE as a speech coder the pitch, voicing and sine-wave amplitudes need to be quantized. It is in the latter operation that IMBE distinguishes itself from MBE and STC. In the latest published version of MBE [55, 56], a set of time-differential sine-wave amplitudes was computed and clustered into frequency bands. The correlation between the residual amplitudes within a band was further reduced using the discrete cosine transform (DCT) and these coefficients were quantized using a non-adaptive bit allocation strategy. Then in [57] a significant performance *improvement* was claimed using an adaptive bit-allocation strategy and the multi-band excitation (MBE) coder became the *improved multiband excitation* (IMBE) coder. Apparently the improved performance of IMBE versus MBE came from tuning the bit-allocation rules to the pitch. This was probably particularly important since the the quantization of the clusters of the residual amplitudes in each of the frequency bands would then be well-matched to the movement of the individual formants. It should also be noted that the published version of IMBE also made use of time-differential coding of the sine-wave phases, and as a result, produced synthetic speech that was reverberant particularly for low-pitched speakers.

It wasn't until the coherent excitation phase model was used that the reverberant quality was eliminated from IMBE system.

It would seem that the IMBE coding scheme would be overly dependent on the pitch and that performance might suffer under channel errors. However, the IMBE system consistently performs well under the most severe channel error conditions and it is apparent that the system is extremely well-designed for robust performance over channels having random and burst errors. This shows that the frequency-domain approach to speech coding using the sine-wave model is amenable to development of effective smoothing algorithms during frames that have been severely damaged due to channel errors and that the IMBE developers have been successful in exploiting this capability. With the impressive performance achieved by the latest version of the IMBE algorithm in the Inmarsat Mini-M tests [53] it is clear that substantial improvements to the quality and robustness of the algorithm have been made over those versions of the system that have been reported in the literature.

7. Conclusions

Since the basic sine-wave analysis/synthesis system can reproduce speech signals with high quality, it provides an ideal basis for the development of a speech coder since, given a high enough data rate, the performance of the codec can be made arbitrarily close to that of the basic system. In fact, using the sine-wave based pitch estimator described in section 3, it is possible to use a harmonic set of sine-waves to produce synthetic speech that is of very high quality. The quality of the coded speech therefore depends on the ability to code the sine-wave amplitudes and phases with good fidelity. Since the focus of the chapter has been to develop speech coders at rates below 4800 b/s, it was necessary to avoid the problem of coding the sine-wave phases and models were developed based on the speech production mechanism that led to the so-called minimum-phase harmonic sine-wave system. Although there is definitely a quality loss when the synthetic phases are used in place of the measured phases, particularly with regard to the replication of sharp voicing transitions, the synthetic speech is of acceptable quality for operation at the lower data rates. Since the minimum-phase harmonic speech coder depends only on the pitch, voicing and sine-wave amplitude parameters, the quality of the low-rate coder depends entirely on the ability to code the sine-wave amplitudes efficiently. The chapter has described one approach that fits an all-pole model to the sine-wave amplitudes and details are given that show how such a system can be quantized for operation at low bit-rates. While the methods described are based on scalar quantization techniques, work is currently underway to explore whether improvements would be possible using vector quantization [59]. It should be obvious that there are numerous methods for coding the sine-wave amplitudes and new techniques are continually being developed, [60]. In fact in the recent pre-selection test for the new DOD Government standard 2400 b/s algorithm [48], five of the eight coders tested, could be classified as sinusoidal coders, while the other three

were waveform coders of the LPC type.

Since the sine-wave parameters provide a frequency-domain decomposition of the speech signal, some of the perceptual properties of the hearing mechanism can more easily be exploited to achieve coding efficiencies. In the development of the amplitude coder, for example, coding gains were achieved by warping the underlying amplitude information on a perceptual scale before fitting the all-pole model. In addition the frequency-domain representation allows for an alternative approach to the design of the synthesis post-filter that, as in other contemporary low-rate coders, is very important in achieving synthetic sine-wave speech that is not muffled. Finally the frequency domain representation provides a convenient basis on which to partition the excitation spectrum into bands so that multi-band voicing decisions can be made that allow for a mixed voicing excitation. Such voicing decisions improve the naturalness of the synthetic speech and increase robustness for speech signals that have been corrupted by additive acoustical noise.

In addition to providing a basis for the development of a parametric vocoder, the sinusoidal model is also being combined with waveform coding methods leading to the class of waveform-interpolation vocoders [11]. This in turn has led to the development of a further decomposition of the sine-wave representation into slowly-varying and rapidly-varying components [12, 13]. By computing the sine-wave parameters at a relatively high data rate ($\approx$ 5 ms), matching the parameters from frame-to-frame, and applying complementary high-pass and low-pass filters to the real and imaginary parts along each of the sine-wave tracks, the rapidly-varying and slowly-varying components of the speech signal can be isolated. If the rapidly-varying components are quantized crudely but often, and the slowly-varying components are quantized accurately but infrequently, high-quality synthetic speech can be obtained at 2400 b/s. This is one of the topics to be discussed in the next chapter.

Acknowledgement

The authors are grateful to Terry Champion, Capt. John Trent, Elliot Singer and Bob Dunn for their support, suggestions, and contributions during the preparation of this chapter.

References

[1] B.S. Atal and J.R. Remde, "A New Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc.*, Paris, France, April 1982, pp. 614-617.

[2] M.R. Schroeder and B.S. Atal, "Code-excited Linear Prediction (CELP): High-quality Speech at Very Low Bit Rates," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc.*, Tampa, FL, 1985, pp. 937-940.

[3] R.J. McAulay and T.F. Quatieri, "Speech Analysis-Synthesis Based on a Sinusoidal Representation," in *IEEE Trans. Acoust., Speech and Signal Proc.*, ASSP-34, (4), 1986, pp. 744–754.

[4] J.L. Flanagan and R.M. Golden, "Phase Vocoder," in *Bell Syst. Tech. J.*, 45, 1966, pp. 1493–1509.

[5] M. Portnoff, "Short-Time Fourier Analysis of Sampled Speech," in *IEEE Trans. Acoust., Speech and Signal Proc.*, ASSP-29, (3), 1981, pp. 364–373.

[6] D. Malah, "Time-Domain Algorithms for Harmonic Bandwidth Reduction and Time Scaling of Spech Signals," in *IEEE Trans. Acoust., Speech and Signal Proc.*, ASSP-27, (2), 1979, pp. 121–133.

[7] P. Hedelin, "A Tone-Oriented Voice-Excited Vocoder," Proc. in *IEEE Int. Conf. Acoust., Speech and Signal Proc.*, Atlanta, GA, 1981, pp. 205-208.

[8] L.B. Almeida and F.M. Silva, "Variable-Frequency Synthesis: An Improved Harmonic Coding Scheme," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc.*, San Diego, CA, 1984, pp. 27.5.1-27.5.4

[9] J.S. Marques and L.B. Almeida, "New Basis Functions for Sinusoidal Decomposition," in *Proc. EUROCON*, Stockholm, Sweden, 1988.

[10] D. Griffin and J.S. Lim, "Multiband Excitation Vocoder," in *IEEE Trans. Acoust., Speech and Signal Proc.*, ASSP-36, (8), 1988, pp. 1223–1235.

[11] W.B. Kleijn, "Encoding Speech Using Prototype Waveforms", in *IEEE Trans. on Speech and Audio Processing*, Vol. 1, No. 4, Oct. 1993, pp. 386-399.

[12] W.B. Kleijn and J. Haagen, "A General Waveform Interpolation Structure For Speech Coding", in *Proc. of EUSIPCO-94*, Edinburgh, Scotland, U.K., Sept. 13-16, 1994, pp. 1665-1668.

[13] W.B. Kleijn and J. Haagen, "A Speech Coder Based On Decomposition Of Characteristic Waveforms", in *Proc. of ICASSP-95*, Detroit, Michigan, May. 16-19, 1995, pp. 508-511.

[14] A. Kondoz, in *Digital Speech: Coding For Low Bit Rate Communication Systems*, J. Wiley, New York, 1994.

[15] R.J.McAulay and T.F. Quatieri, "Low-rate Speech Coding Based on the Sinusoidal Model", *Advances in Speech Signal Processing*, Chapter 6, S. Furui and M.M. Sondhi, Eds., Marcel Dekker, New York, 1992.

[16] L. Rabiner and R. Schafer, *Digital Processing of Speech*, Prentice Hall, Inc., Englewood Cliffs, NJ, 1978.

[17] H. Van Trees, *Detection, Estimation and Modulation Theory, Part I*, Wiley, New York, 1968.

[18] T.F. Quatieri and R.G. Danisewicz, "An Approach to Co-channel Talker Interference Suppression Using a Sinusoidal Model for Speech," in *IEEE Trans. Acoust., Speech and Signal Proc.*, ASSP-38, (1), 1990, pp. 56–69.

[19] T.F. Quatieri and R.J. McAulay, "Peak-to-rms Reduction of Speech Based on a Sinusoidal Mode," in *IEEE Trans. Acoust., Speech and Signal Proc.*, Vol. 39, No. 2, Feb. 1991, pp. 273-288.

[20] A.V. Oppenheim and R.W. Schafer, *Digital Signal Processing*, Prentice Hall, Englewood Cliffs, NJ, 1983

[21] R.J. McAulay and T.F. Quatieri, "Computationally Efficient Sine-wave Synthesis and Its Application to Sinusoidal Transform Coding", in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, New York, N.Y., Apr. 11-14, 1988, pp. 370-373.

[22] T.F. Quatieri and R.J. McAulay, "Speech Transformations Based on a Sinusoidal Representation," in *IEEE Trans. Acoust., Speech and Signal Proc.*, ASSP-34, (6), 1986, pp. 1449–1464.

[23] J. Smith and X. Serra, "PARSHL: An Analysis/Synthesis Program for Non-Harmonic Sounds Based on a Sinusoidal Representation," in *ICMC-1987: Proc. of the Int. Computer Music Conf.*, Computer Music Assoc., 1987, pp. 290-297.

[24] R. Mahler and J.W. Beauchamp, "An Investigation of Vocal Vibratto for Synthesis," in *Applied Acoustics*, U.K., 1989.

[25] R.J. McAulay and T.F. Quatieri, "Pitch Estimation and Voicing Detection Based on a Sinusoidal Model," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc.*, Albuquerque, NM, Apr. 3-6, 1990, pp. 249–252.

[26] W.B. Kleijn, P. Kroon, L. Cellario and D. Sereno, "A 5.85 kb/s CELP Algorithm for Cellular Applications", in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc.*, Minneapolis, MN,

1993, pp. II596–599.
[27] R.P. Lippmann, "An Introduction to Computing with Neural Nets," in *IEEE ASSP Magazine*, 1987, pp. 4–22.
[28] D.B. Paul, "The Spectral Envelope Estimation Vocoder," in *IEEE Trans. on Acoust., Speech and Signal Proc.*, ASSP-29, 1981, pp. 786–794.
[29] M. Unser, A. Adroubi and M. Eden "B-Spline Signal Processing", in *IEEE Trans. on Signal Processing*, Vol. 41, No. 2, Feb. 1993, pp. 821-833. Speech and Signal Proc., Dallas, TX, 1987, pp. 51.3.1.
[30] R.V. Churchill, *Complex Variables and Applications*, McGraw Hill, NY, 1960.
[31] R.J. McAulay and T.F. Quatieri, "Phase Modelling and Its Application To Sinusoidal Transform Coding," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc.*, Tokyo, Japan, 1986, pp. 207–209.
[32] R.J. McAulay and T.F. Quatieri, "Sine-wave Phase Coding at Low Data Rates," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc.*, Toronto, Canada, 1991, pp. 577–580.
[33] R.J. McAulay and T.F. Quatieri, "Multirate Sinusoidal Transform Coding At Rates From 2.4 kb/s To 8 kb/s," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc.*, Dallas, Texas, 1987, pp. 1645-1648.
[34] J. Makhoul, R. Viswanathan, R. Schwartz and A.W.F. Huggins, "A Mixed-Source Model for Speech Compression and Synthesis," in *Proc. IEEE Int. Conf. Acoust.*, Speech and Signal Proc., Tulsa, OK, 1978, pp. 163.
[35] T.F. Quatieri and R.J. McAulay, "Phase Coherence in Speech Reconstruction for Enhancement and Coding Applications," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc.*, Glasgow, Scotland, 1989, pp. 207–209.
[36] J.H. Chen, and A. Gersho, "Real-Time Vector APC Speech Coding at 4800 b/s with Adaptive Postfiltering," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc.*, Dallas, TX, 1987, pp. 51.3.1.
[37] R.J. McAulay, T.M. Parks, T.F. Quatieri and M. Sabin, "Sine-wave Amplitude Coding at Low Data Rates," in *IEEE Workshop on Speech Coding*, Vancouver, B.C., Canada, 1989. Also in *Advances in Speech Coding*, edited by B.S. Atal, V. Cuperman and A. Gersho, Kluwer Academic Publishers, MA, 1991.
[38] R.J. McAulay and T.G. Champion, "Improved Interoperable 2.4 kb/s LPC Using Sinusoidal Transform Coder Techniques", in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Proc.*, Albuquerque, NM, Apr 3-6, 1990, pp. 641-643.
[39] R.J. McAulay, T.F. Quatieri and T.G. Champion, "Sine-wave Amplitude Coding Using High-Order Allpole Models", in *Proc. EUSIPCO-94*, Edinburgh, Scotland, U.K., Sept. 13-16, 1994, pp. 395-398.
[40] R.J. McAulay, "Maximum Likelihood Spectral Estimation and Its Application to Narrowband Speech Coding", in *IEEE Trans. Acoustics, Speech and Signal Processing*, Vol. ASSP-29, No. 1, Feb. 1981, pp. 13-23.
[41] A. El-Jaroudi and J. Makhoul, "Discrete All-Pole Modelling", in *IEEE Trans. on Signal Processing*, Vol. 39, Mo. 2, Feb. 1991, pp. 411-423.
[42] F. Itakura and S. Saito, "A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies," in *Electron. Commun. Japan*, 53-A, 1970, pp. 36.
[43] J.D. Markel and A.H. Gray, in *Linear Prediction of Speech*, Springer-Verlag, NY, 1980.
[44] J.L. Flanagan, in *Speech Analysis, Synthesis and Perception*, Springer-Verlag, New York, 1972.
[45] F.K. Soong and B-H. Juang, "Line Sepctrum Pair (LSP) and Speech Data Compression", in *Proc. IEEE 1984 Conf. Acoust. Speech and Signal Processing*, San Diego, CA, Mar. 19-21, 1984, pp. 1.10.1-1.10.4
[46] E. McLarnon, "A Method for Reducing the Frame Rate of a Channel Vocoder by Using Frame Interpolation," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Proc.*, Washington, D.C., 1978, pp. 458–461.
[47] K.K. Paliwal and B.S. Atal, "Efficient Vector Quantization of LPC Parameters ar 24

Bits/Frame", in *IEEE Trans. on Speech and Audio Processing*, Vol. 1, No. 1, Jan. 1993, pp. 3–14.

[48] M.A. Kohler and L.M. Supplee, "Progress Towards a New Government Standard 2400 bps Voice Coder, in *Proc. Int. Conf. Acoust., Speech and Signal Proc.*, Detroit, Michigan, 1995.

[49] D. Lin, "Statistical Analysis of the BNR Half-Rate MOS Data Set", TIA Speech Codec Working Group, Toronto, 1992.

[50] T.G. Champion, R.J. McAulay and T.F. Quatieri, "Multirate STC and Its Application to Multi-Speaker Conferencing", in *Speech and Audio Coding For Wireless and Network Applications*, edited by B.S. Atal, V. Cuperman and A. Gersho, Kluwer Academic, Boston, 1993, pp. 127-132.

[51] "Inmarsat-M Voiced Codec", in *Thirty-Sixth Inmarsat Council Meeting*, Appendix I, July 1990.

[52] "APCO/NASTD/Fed. Project 25 Vocoder Description", Telecommunications Industry Association Specifications, 1992.

[53] S. Dimolitsas, F. L. Corcoran, C. Ravishankar, R.S. Skaland, and A. Wong. "Evaluation of Voice Codec Performance for the Inmarsat mini-M System" To be Published. in *Proc. 10th International Digital Satellite Conference*, Brighton, England, May 1995.

[54] M. Nishiguchi, J. Matsumoto, R. Wakatsuki and S. Ono, "Vector Quantized MBE With Simplified V/UV Division at 3.0 kb/s", in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Minneapolis, Minnesota, Apr. 27-30, 1993, pp. II.151-154.

[55] J.C. Hardwick, "A 4800 bps Multi-band Excitation Speech Coder", S.M. Thesis, E.E.C.S. Department, M.I.T., May, 1988.

[56] J.C. Hardwick and J.S. Lim, "A 4.8 kb/s Multi-band Excitation Speech Coder", in *Proc. of IEEE Int. Conf. on Acoustics Speec And Signal Processing*, New York, N.Y., Apr. 11-14, 1988, pp 374-377.

[57] J.C. Hardwick and J.S. Lim, "A 4800 bps Improved Multi-band Excitation Speech Coder", in *Proc. of IEEE Workshop on Speech Coding for Telecommunications*, Vancouver, B.C., Canada, Sept. 5-8, 1989.

[58] T.G. Champion and J. Evans, "A Flexible Multirate Speech Coder", in *International Conference on Signal Processing Applications and Technology*, 1993,pp 1440-1443.

[59] R.B. Dunn, R.J. McAulay, T.G. Champion, E. Singer and T.F. Quatieri, "Sine-wave Amplitude Coding Using a Mixed LSP/PARCOR Representation", to be published in *Proc. IEEE Speech Coding Workshop*, Annapolis, MD, Sept.20-22, 1995.

[60] A. Das and A. Gersho, "Variable Dimension Spectral Coding of Speech at 2400bps and Below with Phonetic Classification", to be published in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, Detroit, Michigan, 1995.